Primary data survey: a step-by-step procedure for researchers in social sciences and humanities

Vietkap team

Un-peer-reviewed version 1.2.1

Date: February 8, 2021

Preprint DOI: [10.31219/osf.io/qp9t](https://doi.org/10.31219/osf.io/qp9t)

Chapter 1. Primary data

Preprint DOI: <https://osf.io/f25v7>

February 3, 2021

Abstract

Primary data is inherently the "king" element in the ecosystem of scientific research. Primary data was collected from primary data investigations. This work is not too difficult but requires strenuous efforts and considerable time from scientists, especially young ones with little experience in primary data investigation. Meanwhile, there are merely scant and unsystematic documents on primary data and its investigation methods. In this chapter, definition of primary data, primary data investigation methods and the procedure for data collection by questionnaires will be presented systematically.

Keywords: Keywords: Primary investigation, primary data, questionnaire method, model

Introduction

Data is information about things, events and phenomena, usually in numerical form. In an era of technology where data is deemed "king", the role that data plays has become increasingly important to all industries and fields alike [1,2]. With regard to scientific fields, data plays an even more important role: it is indispensable for research. This is because data contains information on the research subjects. Research papers on how to understand data, how to transform data into useful information, how to identify and solve data problems are derived from data itself. So are ideas, discoveries and innovations. Thus, if scientists have trouble with the ingredient - data

(inadequate or “unclean”, for example), “cooking a decent meal” is out of the question for them.

In an era of information explosion, we have witnessed an exponential increase in data, along with the enhancement in data collection and processing, which is useful for scientists to access data sources [3]. However, this development also gives rise to difficulties in data organization and database design. Moreover, due to convenience and great assistance from technology, scientists are sometimes “lazy” to collect important information, given the considerable effort and cost required, yielding poor research outcomes. Research data is divided into primary and secondary data [4–6]. Since primary data is crucial and its collection is demanding, this chapter delves into the definition and investigation methods of primary data.

Primary data is original and first collected by researchers for the purpose of the study [7]. In each study, scientists sketch out necessary information and data on which to process, analyze and achieve the study's goal. For example, a phone brand plans to launch its latest product and wants to do a survey on the current trends in design and features of a phone. The company will select a group of people that matches the characteristics of its targeted market as a sample and conduct a primary data investigation. Based on the collected data, the company will come up with new directions to meet the general market and reap high profits.

Primary and secondary data are used collectively for highest efficiency. If secondary data is inherited and provides a foundation, primary data will add newness, update and correctness (Table 1). Obviously, primary data is of greater significance than secondary data owing to the following reasons. Firstly, primary data serves only one specific study [7–10], so the information received will perfectly match the study that scientists are implementing. Moreover, information from primary data sources is accurate and complete because it is collected originally and according to specifically designed content.

Table 1. Comparison of primary and secondary data

	Primary data	Secondary data
Advantages	<ul style="list-style-type: none">- Data is "fresh" and is collected directly.- Investigation methods are flexible to suit research questions.- Primary data directly serves current research. More specifically, data collected helps researchers in finding the bottom line of the problem or in predictive analytics.	<ul style="list-style-type: none">- Easy to find as the data is available from previous studies.- The cost of collecting data is cheap, sometimes free, such as data on websites. For example, Statistical Yearbook is a cheap secondary data source.- Ready and suitable (does not take much time analyzing and evaluating).- Secondary data adds value to primary information, especially in clarifying issues, specifying research objectives and choosing primary data, which save time and effort, as well as improve the quality of primary information collection.
Disadvantages	<ul style="list-style-type: none">- High cost.- Time-consuming- Primary data only serves a specific need at a specific time, thus sometimes requiring combining with secondary data investigation.	<ul style="list-style-type: none">- Scales of measurement might be inconsistent with research objectives.- Content might be inappropriate. Previously collected information is inconsistent with research objectives.- Information is outdated and of poor quality.- The data source is not original, thereby decreasing accuracy, which could be attributed to copying and processing the information for other purposes.

As primary data is crucial, its investigation and collection must be done with meticulous attention and precision. In terms of methods, researchers can employ three primary data collection methods: observations, experiments, surveys. Each method has its own advantages and disadvantages, so depending on the topic and field of study, researchers will choose suitable research methods [5,7–14]. To be specific, observations are often used for behavioral science research or abstract research problems that are hard to quantify, experiments for study of natural sciences, medicine and engineering, surveys for quantitative research. We will go into detail about the survey method.

Procedures of primary data collection

The survey method takes a great deal of time and effort to obtain data. Thus, with a view to capturing the most accurate data source, a proper procedure must be complied with. Figure 1 presents the procedure for conducting a complete primary data survey, consisting of the following 5 steps. The first step is to design the questionnaire [15]. The questionnaire is designed based on the research objectives and subject. In the second step, the questionnaire is moved to the focus group in order to be screened and adjusted. The third step is to conduct a small pilot survey [11] to evaluate the questionnaire and to adjust the questionnaire again in order to fit reality. Step 2 and step 3 repeat yet complement each other with a shared view to editing and creating a "perfect" questionnaire before employing it for the next step (step 4). The fourth step is the formal investigation. The final step is to digitalize the collected information, so that we can carry out analysis. The detailed description of each step will be provided in the next chapter.

Figure 1. Procedure for data collection

Besides a proper procedure, human factors are of immense importance in an investigation. A wealth of knowledge and skills is critical for data collectors to create a primary data set that is deemed “qualified”. Of all attributes, honesty is the salient feature in data surveys in general and in primary data investigations in particular. Next is teamwork, management and generalization skills. Given the enormous time and effort involved in primary data investigation, researchers must be motivated and able to motivate others, “keep the fire burning”, so that the collected data is ensured both qualitatively and quantitatively. In addition, communication skills are also

crucial when interacting with interviewees. That the investigator can create a comfortable and relaxed ambience and keep up the volunteer spirit of the interviewees helps to maintain objectivity and accuracy. Moreover, during the investigation, unexpected events or unfavorable situations might arise and disrupt the process, thereby affecting the data source. Hence, problem solving skills are also essential.

To sum up, in this chapter, we have discussed the role and meaning of primary data and illustrated the general procedure of primary data investigation. In the next chapters, main steps involved in a complete investigation will be elucidated.

Review questions:

1. What is data, primary data and secondary data?
2. What are the advantages and disadvantages of primary and secondary data?
3. How many main steps are there in the procedure for a complete primary data set? Name them.

Chapter 2. Questionnaire design

Preprint DOI: <https://osf.io/q3um6>

Date: February 3, 2021

Abstract

In the “ecosystem” of primary data investigation, designing questionnaires is the first and foremost step. This work is not complicated, but often causes a lot of confusion to scientists, especially young ones with little experience in drawing up questionnaires. Meanwhile, documents on designing questionnaires to serve scientific research are sketchy. In this chapter, with a view to solving the aforementioned problem, we will introduce and present the definition of a questionnaire, characteristics of a good questionnaire, the procedure for designing a questionnaire and prerequisite knowledge and skills.

Keywords: Primary survey, primary data, questionnaire method, model

Introduction

In the previous chapter, we have learned about primary data and the procedure for primary data collection. In this chapter, we will delve into questionnaires - a key part as well as the first step in the procedure for primary data investigation [16]. First of all, a questionnaire is simply understood as a set of questions that are closely related and towards a certain topic [8–10,14,15]. In primary data investigation, a questionnaire is employed to collect and store information and data of research subjects. Information in the questionnaire will be digitalized into a valuable data set that will be analyzed by models, statistical formulas to generate new knowledge and reach new conclusions, thereby answering research questions and confirming research hypotheses. Questionnaires are usually classified,

according to methods of conducting, into face-to-face questionnaires, mail questionnaires and online questionnaires [5,7,11]. Of the three, face-to-face questionnaires are most frequently used. Nowadays, in a technology-driven era, online questionnaires have become more diverse and are gaining popularity. Overall, the main contents of all three methods are the same, with only a minor difference in style and ways of questioning. Details of the two key questioning methods are shown in Table 2

Table 2. Comparison of questionnaire methods

	Face-to-face	Online
Characteristics	- Researchers meet the interviewees in person to fill the questionnaire.	- The questionnaire is designed to present itself to interviewees via websites or electronic devices located in crowded places (supermarkets, shopping centers, etc.)
Advantages	<ul style="list-style-type: none"> - Access a full range of research subjects, regardless of educational background, age, income level and occupation, which is critical to social research. - Effective when done in a particular geographical area with favorable traffic conditions. - Researchers are in a good condition to explain the purpose of the research and the meaning of the questions, thereby reducing inappropriate answers. 	<ul style="list-style-type: none"> - Large-scale and can be conducted and digitalized quickly and effortlessly as information is automatically saved on the spot. - Thanks to software and apps, the questionnaire can reach a wide audience.
Disadvantages	<ul style="list-style-type: none"> - Requires a lot of time, effort and money. - For research that requires a cornucopia of response on a large scale, there would be many difficulties. 	<ul style="list-style-type: none"> - Initial cost (building software, web, app or buying electronics, et.) can be quite high. - Cannot reach underprivileged groups.

	<ul style="list-style-type: none"> - Results depend a lot on the subjectivity of both questioners and the questioned, so certain deviations exist. 	<ul style="list-style-type: none"> - The research purpose and the question content might be misinterpreted as it depends on the mindset and perceptions of the questioned; deviations and noise may exist. This is analogous to a disadvantage of the mail questionnaire method.
Conclusions	<p>In general, all methods have their own merits and demerits. Thus, researchers need to utilize each method flexibly to achieve efficiency. Of all three, face-to-face questionnaires should be given number one priority concerning accuracy and objectivity - the top criteria in scientific research, and accessibility as well. The remaining two methods can be used flexibly depending on the research purpose. For uncomplicated research with transparent questions and on a focus group, they are recommended given their advantages of time, effort and cost saving</p>	

Questionnaire structure

In terms of structure, a questionnaire usually has three parts: an introduction, a body and an end.

- *First:* the introduction. It should clearly state the research purpose, the project title and the survey research unit. A successful introduction containing well-chosen images and logos serves as a hook to capture the respondents' interest. Moreover, we can add some more information such as the reasons for participation and put an emphasis on the role of the interviewees in order to motivate them. Below is an example of the introduction of the questionnaire on land use of households living in some mountainous communes of Nghe An province [17], which clearly states the survey research unit, ID of the questionnaire (making it easier to categorize), illustrations, general layout, acknowledgments and guarantees (Figure. 2).

ID:

Farmers' livelihood strategies and perceived constraints from the poor and non-poor households: A dataset from field survey in Nghe An, Vietnam

Structure of the questionnaire		Number of questions	Pages
Part A	Household land use structure	9	2
Part B	Household income structure	15	3
Part C	Perceived factors hindering production forest planting	23	4
Part D	Household information	15	7
Total pages			8

Figure 2. A snapshot of a page cover of the questionnaire.

- *Second:* the body - the core of a questionnaire, which consists of questions bearing on the research topic. The content should be valid, complete and succinct. On top of that, questions should be arranged in a logical and interesting way in that the overwhelming number of questions in the previous section could discourage respondents. For example, in the study “Understanding the satisfaction of citizens with governments’ response to COVID-19”, the body is divided into different question categories

corresponding to each research question. These questions are presented in a general-to-specific sequence. The questionnaire includes Likert scale¹ questions that help to gain basic yet useful information that shows whether the interviewees were properly aware of the survey, or to apply nominal scale in identifying suitable research subjects. These questions will acquaint interviewees with the study, assisting with the response to the ensuing part. Regarding format, they are an affirmation with different choices of certainty (usually from 1 to 5). Next are questions about public perceptions of COVID-19 impacts and attitudes towards preventive measures. The last important part of the body is questions about voluntary payments, contributions to social organizations.

- *Third:* the end. This section consists of two small parts: personal information and acknowledgment, which must be kept at the end. Personal information will reveal name, age, occupation, hometown, demographics and so on. However, this section needs to exclude intimate information such as phone number or Gmail, for it might worry and discourage respondents from answering the next parts. Lastly, the acknowledgment should ensure succinctness and the designers' sincerity.

Characteristics of a good questionnaire

A good questionnaire should be:

- Complete, succinct: A questionnaire must cover all the information that researchers need to collect, meaning it is "complete". Succinctness is important for two reasons. Firstly, a pithy questionnaire will avoid potential errors and misunderstandings by the questioners as well as the questioned. If a question is unduly long, it causes confusion. Also, the more questions participants need to answer, the more fatigue and boredom they might

¹ A question that uses a 5 or 7-point scale, sometimes referred to as a satisfaction scale, that ranges from one extreme attitude to another. Typically, the Likert survey question includes a moderate or neutral option in its scale.

endure. All of these will exert an adverse influence on the accuracy of the collected information. Secondly, a succinct questionnaire is time- and cost-saving for the survey.

- Clear, easy to understand and answer: Given the diverse backgrounds of the interviewees, transparent questions are necessary to minimize errors and reduce the need for further explanation from the questioners.
- Digitizable: This feature enables in-depth analysis, thereby retrieving scientific information. Without it, the questionnaire would only bring about invalid, illogical and unreasonable conclusions.

The procedure for designing questionnaires

Designing a quality questionnaire is a prerequisite for the success of a scientific study. The foundation for a questionnaire is the research purpose and topic. In addition, the questions should closely and directly bear on the research hypothesis or the research model (Figure.2). The designed questionnaire also needs to suit the research subject. Answering questions about information like the age range or occupation of respondents would help design a suitable questionnaire to create excitement and eagerness. It is also essential to note the geographic location in the questionnaire. For example, in areas where ethnic minorities with low literacy rate inhabit, researchers need to design a questionnaire that is concise and very easy to understand.

Figure 2. Relationship between the variables and the questions

Based on intimate knowledge and empirical experience in designing questionnaires, we propose a 6-step procedure to draw up a questionnaire as follows:

- Step 1. Construct the research hypothesis and the purpose of the topic
- Step 2. Learn about the research models that were used to find important variables. In previous studies, was a linear regression model utilized? How many relevant variables were there?
- Step 3. Build a model related to the variables. For example, in the survey of land use in upland communes of Nghe An province, we constructed a multivariate linear regression model where the dependent variable is income and the independent variables are factors affecting the household income such as land use, loan, labor experience.

- Step 4. Check and complete the model according to actual conditions via preceding survey. After this survey, add some important variables or remove unnecessary ones from the model.
- Step 5. Develop a detailed research outline and draw up a questionnaire. In this step, we will design a detailed questionnaire. Each question is intended to collect data according to each point in the detailed outline and may include variables. For example, in the model, the education level of the head of household is a variable. Thus, a question to obtain this type of information is designed.
- Step 6. Check the information and complete the questionnaire.

Question types and data types

In order to design a questionnaire, questions are divided into close-ended and open-ended questions. A close-ended question is a question with predefined answers, and the respondents can only choose between them [7,11]. Examples include dichotomous questions (yes/no questions), Likert scale questions (measuring importance, satisfaction level, etc.), multiple choice questions. An open-ended question is a question that respondents give their own answers (about their age, hometown, occupation, etc.).

Irrespective of question types, collected answers must fall into one of the following scales of measure: nominal, ordinal, interval and ratio. Questionnaire designers must understand what data type of each question is because it determines coding and analytical methods.

Question types and data types

- Nominal scale: It is used to classify subjects according to different features. For example, according to gender, “male” can be encoded as 0, “female” as 1; or to find out if the household is poor or not, “yes” is encoded as 1 and “no” as 0.

- Ordinal scale: It is used for ranking or comparison according to a certain criterion. Take the classification of students' academic performance as an example: 1 = Excellent, 2 = Good, 3 = Average, 4 = Fair, 5 = Poor.
- Interval scale: It is used for questions with quantitative answers. This scale has evenly spaced and continuous intervals. For example, questions with the answers denoting: 1 = totally disagree; 2 = disagree; 3 = neutral; 4 = agree; 5 = totally agree.
- Ratio scale: It is also used for questions with quantitative answers. However, this scale is different from the interval scale in that it contains information representing precise quantity. For example, person A is 30 years old and earns a monthly income of 20 million VND. The information about income could be used to compare with person B, whose monthly income is 10 million, half of that of person A.

Scales of measurement act as guidelines for questionnaire designers to know if their questions are “correct”. A “correct” question will yield answers that can be measured and encoded into one of the four scales above.

The quality of the questionnaire determines that of the research. Therefore, it is crucial to enhance the questionnaire. To achieve this end, researchers could consult experts who are highly experienced in designing questionnaires or participate in primary data investigations. Another way is through the focus group. Thanks to a pilot survey, researchers take a holistic approach to factors affecting the quality of the information obtained, analyze data to test if the results meet expectations, thereby adjusting the questionnaire accordingly. Moreover, listening to and synthesizing opinions of surveyors about difficulties that they face, for example, and of the surveyed people about the questions, unnecessary or inadequate, will also help improve the quality of the questionnaire.

Conclusion

Although the procedure, structure and content of the questionnaire are not

too complicated, profound knowledge and a set of skills are required of a questionnaire designer. Each and every question in the questionnaire must have a clear theoretical basis and purpose. For young researchers with little experience in primary data investigation through questionnaires, especially students who start learning to “do science”, participating in training classes and designing a questionnaire first-hand would stand them in good stead. This learning along with practicing will help them draw experience in designing questionnaires for the next time.

Chapter 3. Focus group

Preprint DOI: <https://osf.io/nfczd>

Date: February 3, 2021

Abstract

A focus group is a component of the "ecosystem" of primary data investigation. The mission of a focus group is to test and enhance the questionnaire. Up to the present, "focus group" has been introduced in some scientific research methodology documents. However, this content is presented mainly in a theoretical manner, with little or no real-life examples. In addition, it is not succinct enough for readers, especially young scientists with little experience in forming and promoting "focus group", to fully understand. In this chapter, the definition of a focus group is presented systematically, being backed up by a real-life example. Next are the procedure and notes on focus group formation. Finally, we summarize some prerequisite skills to found and operate a focus group.

Keywords: Qualitative research, questionnaires, opinions, teamwork skills

Introduction

In the 5-step procedure to conduct primary data investigation, designing a questionnaire is followed by utilizing the questionnaire for test and completion with the aid of a focus group [16,18]. A focus group is a group of around 5 to 10 selected individuals who voluntarily participate in a discussion on a survey topic [11,14,19]. The researcher selects and gathers these members in order to evaluate, edit, supplement and complete the questionnaire. Participants are encouraged to voice and share their views on the questionnaire. The result of this step is a relatively complete questionnaire that can be used for the next step. To execute a focus group

well, researchers need to fully understand the work entailed and the requirements for each task (Figure 3).

Figure. 3. Focus group, source [20]

Procedure

The work of a focus group survey consists of three chronological steps: before, during, and after a focus group survey.

Step 1. *Before*. There are three things a researcher needs to do after having a "satisfied" questionnaire. First, to summarize and get the gist of the questionnaire, grasp the objectives and meaning of each question group, related concepts, terms, and pay heed to uncertain parts. This would enable a more detailed discussion with the focus group, yielding a "best" questionnaire. Second, to select members of the focus group according to some criteria consistent with the content of the questionnaire. They should

show critical thinking, enthusiasm, sociability and willingness to contribute ideas. Third, to estimate the number of participants (from 5 to 10) and decide the place of meeting that meets the requirements of privacy, quietness and comfort.

Step 2. *During.* The onus falls upon the researcher to lead and motivate the participants to voice their opinions, which is the determining factor to the success of the discussion. At the beginning phase, the coordinator presents an overview of the purpose of the study, the questionnaire content and explains related terms. Then, the group will review the questionnaire and discuss. That a relaxed ambience prevails will encourage contribution. The coordinator is responsible for taking notes of useful opinions, keeping the discussion on track, avoiding rambling, thus ensuring efficiency.

Step 3. *After.* The researcher relies on the notes synthesizing the participants' opinions to screen and retrieve useful information. Then, these contributions will help to complete the questionnaire by removing redundant, repetitive questions and adding appropriate ones.

Following the three steps above, the researcher brings teamwork into play to enhance and complete the questionnaire that will be used in the next step of pilot investigation.

Real-life example

At the end of 2020, the Vietkap team carried out a survey whose title is "Understanding the satisfaction of citizens with governments' response to COVID-19". The team applied focus group to revise and complete the questionnaire as follows:

- Step 1. In this step, synthesizing and generalizing also helps the author group to go over the questionnaire. Then, a researcher was selected to run a focus group in step 2. Our author group consists of 5 people with extensive knowledge and intimate experience in primary data investigation. The preliminary questionnaire was sent to them beforehand to examine and voice

their opinions during a group meeting. An evaluation of each and every question was made to rank levels of agreement in descending order as follows: “completely agree and not adjust”, "agree but adjust", "disagree and replace” and "delete". After editing the questionnaire accordingly, the author group reached a unanimous decision on a focus group of 10 students gathering in a spacious and quiet meeting room.

- Step 2. Manage the focus group: Before the meeting, selected students were informed of the plan and sent the questionnaire. During the meeting, the representative of the author group presented key information about the research topic and basic concepts in the questionnaire. Then, the students took a trial interview and assessed the questionnaire’s length, difficulty levels and appropriateness. Spelling errors and confusing jargons are also meticulously checked and corrected. The results of the meeting lived up to expectations and met the goals of revising and enhancing the questionnaire, some questions being removed, some added. At the end of the discussion, the moderator summarized the changes that the majority consented.

- Step 3. The gathered information from the meeting was synthesized and edited by the author group. Results showed that 20 out of 100 questions required adjustments to become more explicable, 5 added, 13 removed, and some word choice errors and typos were corrected. It could be concluded that the focus group achieved high efficiency and accomplished its mission. As a result, the author group obtained a questionnaire that was deemed "satisfied" enough to move to the pilot survey step.

Prerequisite skills

In the process of forming and running a focus group, the team leader plays a vital role, responsible for planning, selecting and connecting members, leading and navigating the discussion. In reality, managing a focus group encounters several difficulties due to a wide diversity of personalities and perspectives. It is never easy to reconcile differences and reach a consensus in a group. Hence, finesse and persuading skills are required of a

coordinator. Besides, we should not ignore the role of team members whose constructive criticism and positive support are decisive to the success of the group.

Focus group operation requires flexible application of many skills. Right from the first step, on inviting volunteers to join the focus group, persuasive skills are essential. Afterwards, a plan to inform, send questionnaires to members and arrange a meeting needs to be drawn up, requiring planning skills. In the group meeting, good presentation and listening skills mean that the scientist can raise and explain the issue, stimulate discussion and summarize ideas. Bear in mind that during a discussion, many unexpected events might arise, calling for problem-solving skills. All in all, in order to form and run a focus group effectively, scientists must develop a set of key skills.

Conclusion

In the ecosystem of primary data investigation, focus group is a crucial qualitative method. Drawing on our own experience, we realize that if “focus group” is well managed, the questionnaire will improve markedly, serving both pilot and official surveys. Great efforts and academic knowledge are inadequate to run a focus group. Scientists, especially young ones, also need to equip themselves with a set of key skills. Of all, teamwork and problem-solving skills are of pivotal significance.

Review questions

1. Define a focus group? What role does it play in a primary data investigation?
2. Present the procedure for creating and operating a focus group.
3. What skills are required to operate a focus group effectively?

Chapter 4. Sampling

Preprint DOI: <https://osf.io/w9qks>

Date: February 3, 2021

Abstract

In previous chapters, we have understood the importance of primary data in scientific research and the procedure for conducting a primary data investigation. After completing a questionnaire, the next step is to conduct a field research. In this chapter, we will clarify two questions: who are the respondents and are the number large enough? Because of limited resources and time, most studies were unable to collect data on the whole population of study subjects but on selected participants. This work is called “*Sampling*”. Based on the analytical results calculated from the investigated sample, researchers can extrapolate scientific information and conclusions about the whole population. The sampling process consists of 2 main tasks: selecting the target population and deciding on the number of samples to investigate (sample size). Currently, there are many scientific documents on sampling, but this content is complex because it is related to knowledge of probability, statistics, and contains many concepts, formulas. Consequently, readers, especially young researchers, often find it difficult to understand and to apply. In this chapter, we do not cover all sampling methods as well as the complex formulas, yet the main goal is to present the right, basic knowledge in an apparent and easy to apply way. For more advanced knowledge, readers can look at references.

Keywords: Study population, sample selection, sample size, sampling

Introduction

Unlike studies without the purpose of generalizing from a sample, those that try to extrapolate and generalize about characteristics of the entire population from those of a sample will depend greatly on sampling (Figure 1). Therefore, in the primary data investigation, sampling is considered the overarching step [5,7,16,18]. We will start this chapter with the most basic concepts of research population, sampling frame and sample size. Next, we will delve into sampling methods and calculate sample size using examples of research in social sciences (Figure 4).

Figure 4. Sampling. source [11]

Population, sampling frame and sample

A ***study population*** is a collection of all research subjects or individuals possessing characteristics that meet the criteria of the research, from which a sample is drawn. A ***sample*** is a subset of the research population including a number of research subjects, selected according to a certain rule and usually representative of the research population [7,11].

A ***sampling frame*** is simply understood as a collection of individuals in the target population [11]. After identifying the research population, a sampling frame needs to be developed for data collection. In this step, scientists need to avoid three types of “false” frames. First is a frame that contains too many individuals, some of which are even not in the target population. Next is one that contains too few individuals that is either in or out of the target population. The third “false” frame contains an incorrect set of individuals or is itself not part of the target population. In order to frame a sample, scientists need to determine individuals in the target population and the sample size. For example, in the study of students’ academic performance, the frame could be the list of students in separate grades, or according to performance, another the list of teachers. Depending on the research objective or questions, scientists build a suitable framework from which an investigation could be conducted.

Sample size is the number of individuals of a sample selected from the population to be surveyed [5,7]. Thus, a sample could be understood as a subset of the population and sample size is the size of this subset. For example, a research on income of residents in a province has the whole population of one million. The researcher decides to conduct a survey on 100 thousand people belonging to various income groups, then 100 thousand people is the sample size. The scientific basis for selecting the sample size is the Law of large numbers in probability and statistics. This rule states that as a sample size grows, its mean gets closer to the average of the whole population.

Sampling methods

There are two sampling methods used by scientists, including probability sampling (random) and non-probability (non-random) sampling.

Probability sampling involves completely random selection, irrespective of subjectivity. As a result, the probability of each observation being selected is similar, thereby likely to select a sample representative of the whole population. Since the sampling error can be calculated, we can apply statistical estimation methods, test statistical hypotheses, and process data to extrapolate results from the sample. However, this method has some limitations. First, it is difficult to apply in that a specific list of the general population could not be identified. Besides, it demands a plethora of time, cost and human resources for data collection if the subjects are scattered over a geographic area.

Non-probability sampling involves selecting a unit of observation depending on the subjectivity of the chosen. This method is often used when the random sampling method is unviable, such as in novel surveys without any information about the subjects, or subjects dispersing, volatile or belonging to disparate groups. Non-random sampling is not completely based on mathematics but requires a close combination between theoretical analysis and social practice. Hence, it, to a large extent, represents the subjectivity of the selector.

Determining sample size

In sampling, determining sample size always causes confusion for researchers, even for experienced ones. Common problems are in deciding on the size and in explaining why that number: 300 and not 350 or 500, for example. In this section, we will discuss two approaches to determining sample size. The first is to use mathematical formulas, and the second, for some social studies, we can employ a combination of "expert, experience and mathematical formulas".

Using formulas: In practice, the entire population can be divided into several groups with different characteristics, from which there are different ways to calculate the sample size. Table 3 shows two simplest and most common formulas for sample size determination. Besides them, there are many other techniques using statistical analysis models to calculate the sample size like the Exploratory Factor Analysis (EFA) or regression model. However, in this chapter, we will not go into details to avoid complication.

Table 3. Sample size calculation formulas

	Whole population size unknown	Whole population size known
Formulas	$n = \frac{Z^2 p(1-p)}{e^2}$	$n = \frac{N}{1 + Ne^2}$
	n: sample size Z: confidence level (usually at 95%), standard value of 1.96 p: probability e: error	n: sample size N: whole population size e: error

From the formulas in Table 3, it can be seen that sample size calculation depends on many factors. However, given that representative factors are the salient point, there are four main elements that influence the determination of the sample size.

- First is reliability of the data. That is, the degree of certainty and generalization for the whole. The more generalizable the data is, the more reliable it is.

- Second is small and acceptable errors. Although errors are inevitable, we must limit and minimize them to ensure accuracy for estimate of the sample.
- Third is the types of test statistics, some of which require a minimum sample size to have meanings.
- Fourth is the size of the whole population. The ratio of sample to population must meet a certain requirement. For example, the sample size to be surveyed for a city of 5 million inhabitants would be larger than that for a commune of 5,000 inhabitants.

Determining sample size in social science research: Although sample size determination is not too complex, scientists with many years of research might face several difficulties in identifying reasonable ones for their studies. The research on factors affecting household income could be cited as an example. In social science research, scientists often use regression models to process data. Thus, the sample size is also calculated in order to serve this purpose. Drawing from experience in primary data investigation, we noted that there are three ways to determine the sample size.

- The first is that researchers can rely on the “expert” method - consult experienced scientists for advice. This method is fast and effective if the right specialist is found.
- Second, researchers can apply the rule of thumb in determining sample size. In other words, a sample size ranging from 300 to 500 is acceptable [5].
- Third, scientists can use different levels of confidence and error to construct a table of minimum sample sizes according to distinctive error and confidence scenarios. From that, a sample size should have a minimum of 271 items to ensure a 90% confidence level or more, with 5% error. When the sample number is 385 or higher, the confidence level will be 95% or more [8]. For higher reliability, scientists should determine a sample size of 400 or more. The reason is that after

the survey, they have to discard a certain number of faulty observations or outliers, usually accounting for 3-5% of the total sample size.

In summary, sampling is a critical step in the primary survey “ecosystem”. A bigger sample size means that the characteristics of the population is clearly presented. However, it also means higher cost and more time. In contrast, if the sample size is too small, it will lack objectivity and show no difference between the subjects. With the goal of honing this step, scientists need to firmly grasp the research objectives, the Law of large numbers and the theory of statistical probability. A suitable sample size is neither too big nor too small, meeting the purpose and status of the study. Flexibility in determining the sample size according to financial and human resources should also be heeded. However, under any circumstances, the sample size should meet at least one criterion according to the formula for sample size calculation or be based on the sample size of similar conducted studies.

Review questions

1. Define sample and sample size. Give examples?
2. What is the scientific basis for sampling? Which factors affect sampling? Why?
3. Present the formula for sample size determination and give examples?
4. Present sample selection methods. Give examples and describe the advantages and disadvantages of each method?

Chapter 5. Pilot survey

Preprint DOI: <https://osf.io/dwhja>

Date: January 31, 2021

Abstract

Pilot survey is the next important step in the primary data investigation ecosystem. The most popular mistakes committed by young and inexperienced scientists are underestimating the importance of the pilot survey, thereby failing to devote a sufficient amount of time and focus to this stage. In fact, however sophisticatedly and carefully planned a survey is, limitations and flaws are inevitable in actual practice. Experimental study facilitates scientists in detecting incurred problems and promptly fixing them before official survey, which improves research quality and saves time, financial and human resources. Even though there are manuals touching on this aspect, the content is scanty and lacking in examples to further understanding of every single step. In this program, we will discuss in details procedures and contents in pilot survey.

Keywords: Questionnaire, focus group, validity, reliability

Introduction

Pilot survey is a preliminary investigation conducted on a small scale, and its sample size [21] is 5-10% of the planned population. The main objective is to test and upgrade questionnaires [16,18], for rarely is a questionnaire suitable for large-scale investigation right from the beginning. After initial changes by means of investigation into the focus group [22], the questionnaires are

continuously subject to change under experimental investigation until researchers receive multi-dimensional information from surveyors and interviewers and produce the best version for official survey. To further understand, we need to master three areas: (i) participation selection, (ii) required output of experimental investigation (objective), (iii) monitoring of pilot survey.

Participants

When making a list of participants, in most cases, scientists will do extensive research and select experienced ones. This serves to reduce potential risks to a minimum in the processes of both experimental and official investigation, thereby saving time and efforts for research instructions. However, reality shows that these expectations are hard to live up to when it comes to large scale surveys, which compel researchers to offer detailed instructions to participants. While one or two discussions with researchers are sufficient for the experienced ones to have a firm grasp of the questionnaire and clear up all confusion, their inexperienced counterparts have to spend a longer period of time on training and experience sharing during the investigation.

Objectives of pilot survey

The main purpose of the pilot survey is to improve standards of the questionnaire. A questionnaire of good quality is composed of three criteria including validity, reliability and effectiveness [5,7,11].

Firstly, the validity of a questionnaire needs checking. A questionnaire is set up by a small group of experts within a short period of time, so it is inevitable that there remain errors in wording, spelling or overlap among questions or

exclusion of crucial questions. A pilot survey reaches a number of people with high response rates prove the validity of the questionnaire

Secondly, the reliability of a questionnaire is assessed via participants' responses. Researchers invariably aim for the most objective answers with a high degree of reliability to enhance research practicality and accuracy. To accomplish this, questions with precise wordings and presented in a succinct manner is of great importance, minimizing any potential confusion or misunderstanding. Field interviewers can test the reliability of the questionnaire by asking for the interviewees' perspective of the question. Concretely, should there exist a considerable number of people whose inference strays far from the original intention, then there are underlying issues misleading readers.

Thirdly, a questionnaire is considered effective if it narrows down questions and reduces asking time to a minimum and amasses as much useful information as possible. Effectiveness is assessed by embracing field interviewers' and participants' responses as to which questions, in their opinions, are either difficult to answer or not practical, or which groups of questions overlap with one another and thus can be combined together, etc. Some suggested questions to assess the validity, reliability [5,7] and effectiveness are as follows (Table 4).

Table 4. Some questions to test questionnaire

Criteria	Questions
Validity	<p>1. Are there any flaws in the questionnaire? Flaws come in various forms such as spelling and wording. Reality demonstrates that many questions are clear in writing but inappropriate for interviews.</p> <p>2. Is the questionnaire logical? Separate questions may be accurate, but improper arrangements are likely to cause confusion in the survey process and incur errors. Therefore, it is advisable that questions related to one aspect under study appear adjacent to one another.</p> <p>3. Do questions contradict real situations? There are questions that, albeit lies within the researcher's plan, bear no practical value. These are conducive to a loss of interest in interviewees and thus should be omitted.</p>
Reliability	<p>4. Do interviewers fully understand the questions? Complex wording in a question can be misleading or incomprehensible to interviewers.</p> <p>5. Do participants fully understand the questions?</p> <p>6. Are the questions ambiguous?</p> <p>7. Do the interviews attract respondents' attention and intrigue them throughout the survey process? If their attention is distracted, then the quality of data can be open to question.</p>
Effectiveness	<p>8. Are respondents capable of answering questions? It is obligatory for the questionnaire to eliminate time-consuming or complicated questions and to ensure that respondents can provide answers at ease.</p> <p>9. Are given answers to closed-ended questions provided sufficiently? Provided options can be insufficient and out of touch with reality, so field interviewers should take notes to help researchers acknowledge and supplement the questionnaire.</p> <p>10. Are provided options clear enough to distinguish one from another? If most of the respondents give the same answer, the options may need to be revised to make it distinguishable.</p> <p>11. Are questions and answers brief and concise?</p>

Monitoring

Participants should be offered careful supervision while a comfortable atmosphere is maintained during the survey process. Furthermore, pilot survey affords field interviewers scope for becoming fully conversant with the questionnaire and their assigned tasks. Researchers are required to firmly grasp the plan and roadmaps of members, check and listen to members' feedback on a regular basis to make prompt adjustments. Moreover, researchers are supposed to remind members to take notes of gathered information (besides the questionnaire), which serves as a valuable overview so that researchers themselves can make proper changes to the questionnaire as well as their own perspectives in their research.

Another important issue that needs managing is operating costs. The operating costs are heavily reliant on the scale of the survey, and how the group was run when it was tested. Optimizing financial resources while still yielding desirable results in the survey is the top priority of the scientists. All costs, including estimated and incurred ones, are retained for careful assessment of the appropriateness and efficiency of capital. The accurate assessment of costs in the pilot survey stage will assist scientists in estimating the costs of official surveys and devise an exhaustive plan.

Handling problems arising before, during and after the pilot survey

In summary, the pilot survey is a rehearsal for the official survey, which is an important step in the primary investigation ecosystem. The fundamentals of the pilot survey is to check the operational quality of the survey team as a whole, check the quality of the questionnaire and evaluate costs. Before carrying out the pilot survey, meticulous, scientific and rational planning is of

paramount importance. In operation of the pilot survey, organizers' flexibility in dealing with arising problems plays a decisive role. After the pilot survey, it is imperative to review each point for errors and make improvements. Although the pilot survey is essential, with a familiar research topic and small-scale research, scientists with a wealth of knowledge, skills and experience in conducting primary investigations can skip this stage to cut costs. For young scientists, trial investigation should be considered a mandatory stage, for it is an opportunity for them to practice and accumulate a wide array of indispensable skills such as organizational skills, management skills and problem-solving skills.

Review questions

1. Present the main contents of pilot survey
2. Name the questions to exam the validity, reliability, and effectiveness of questionnaires

Chapter 6. Final survey

Preprint DOI: 10.31219/osf.io/2arhj

Date: February 3, 2021

Abstract

A pilot survey is conducted to test a survey design and increase its likelihood of success, followed by a final survey - the most crucial step in the primary data investigation ecosystem. Its objective is to obtain standard information and or data for in-depth research. What poses questions is how to conduct final survey successfully? Well handling this work is fairly challenging, especially when it comes to young scientists of insufficient experience in primary data investigation. This chapter will present a winning formula for final survey as a source of useful reference for scientists. Procedures, content and principles constitute this formula in primary investigation.

Keywords: primary data, questionnaire, operation, procedure, principles

Introduction

Compared with the pilot survey [7,23], a final survey is conducted on a large-scale with a much higher degree of complexity, and its duration depends on research types, characteristics and conditions [5]. However, final survey usually lasts for one or two weeks, which is believed to be a period of optimal productivity and quality of the research group. This chapter is divided into three main parts: (1) Investigation contents associated with separate steps in the main study, (2) a list of tasks for check before, during and after the field investigation. (3) Principles of reacting to potential risks/situations arising in the operation process (Figure 5).

Figure 5. A three-dimension diagram of final survey

Procedures: Preparation - Operation - Evaluation/Synthesis

The 3-step procedure involves three phrases: before, during and after field data investigation. Pre-investigation can be conducted either at the office or at home, followed by on-site work and comprehensive evaluation at the office in phase 2 and 3, respectively.

Phase 1: Preparation

1. *Planning.* In this phase, scientists collaborate with the research team to devise a concrete plan for official investigation in terms of questionnaires, survey sample determination, cost estimates, human resources, material resources and expected outputs.

2. *Communicating with locals.* Scientists will search for possible channels to contact locals including commune and village officials. This phase aims to

inform, inquire and hold preliminary discussions with locals about the upcoming empirical survey to drum up support and coordination to work effectively and efficiently.

3. *Preparing questionnaires.* This phase requires scientists to complete the questionnaire design. Throughout the pilot survey, the revision process is ongoing until the questionnaire is finalized for use in the main study.

4. *Preparing human, material and financial resources.* Scientists will provide precise estimates for labor within budget and the time limit. Besides, financial resources, facilities and tools need working out to serve the official survey process. Step 4 demands elaborate preparation in readiness for investigation.

5. *Plan approval.* Scientists will recontact the locals for confirmation as to the investigation duration, scale and content. This step aims at locals' willingness to participate in the final survey.

Phase 2: Operation

6. *Liaison with the communal authorities.* In this phase, the survey crew will move to local communes, and usually meet up with the authorities. Secondary documents about the general socio-economic status of each commune are expected to be collected. Another vital output is the referral to higher authorities of districts and hamlets prior the crew's arrival. In many cases, the authority will provide accommodation for the survey crew.

7. *Liaison with hamlet authorities.* A key component of this phase is communication and liaison with the hamlet leaders. The outcome should be a list of households who are eligible to join the official research, with food and shelters provided.

8. *Interviewing.* In this phase, the research crew will begin the interview process with households. Ideally, district authorities will accompany the crew in order to boost the research's legitimacy and reliability, as well as to solve

any arising issues. The expected outcomes of this phase are correctly and detailly filled out questionnaire forms.

9. *Finishing.* The investigation concludes. The survey crew organize a farewell meeting with the authorities, then safely travels back. Upon returning home, the most pressing task is finishing, checking, and inputting data from filled-in questionnaires to the datasheet. Usually this task should be completed promptly within the first week of returning. In order to succeed, the research crew ought to prepare input spreadsheets in advance, and devise a thorough plan on questionnaire checking.

Phase 3: Comprehensive evaluation

10. *Summary and evaluation.* A meeting will be held for evaluation of results of primary data investigation. Members raise their voice as to all aspects ranging from questionnaires to human resources, finance and operation methods to draw valuable lessons for future further research.

Worklists

Details of the list of tasks are shown in the Table 5:

Table 5. A list of 11 work contents in the final survey

No.	Categories	Details
1	Questionnaire	Hard copies Soft copies for printing nearby survey areas (Yes/No)
2	Human resources	Interviewers (experts, collaborators) Local guides (hamlets, communes) Interviewees (locals)
3	Vehicles	Rental (car, airplane, train, motorbike) Self-prepared (Yes/No)
4	Tools	Laptop A4, A0 paper Ballpoint pens, markers Map of terrains, land usage Board / notes... Glue, tape
5	Logistics	Meals (Yes/No) Backup food (Yes/No) Accommodation at research areas (hotels, local houses) Gifts for supporting officials and guides (Yes/No) Gifts for interviewees (Yes/No) Gifts for house owners (Yes/No)
6	Medicine	Cold, flu, and fever medicine Medicine for diarrhea Insect repellents First aid kit (Bandage)
7	Means of communication	Cellular data availability Internet availability Phone top-ups (Yes/No)
10	Administrative papers	Referral letter (Yes/No) Verbal introduction (Yes/No)
11	Expenses	Cash; Cash account Money for local specialties and souvenirs (Yes/No) Supplementary budget

Principles of reacting to potential risks/situations that arise

Despite meticulous preparations for main studies, scientists still encounter various situations unaccounted for in local areas. There are 4 main types of incurrence. The first type is increased workload. Scientists may face unwanted issues with regards to the actual work, namely a failure to meet with the locals or interviewees, or a member of the researching group having health problems or having to leave the crew due to personal matters. The second issue lies in the weather, which is also an external factor, involving but is not limited to extreme heat, heavy rain, thunderstorms, blizzard... Thirdly, time may also act as an impediment. This is mainly subject to either subjective or objective reasons, which can lead to lengthy interviews that stretch past planned schedules or force the crew to extend the proposed timeline in order to manage the workload successfully. The final and also most commonly occurring issue is arising fees. These incurred costs may stem from transportation, or compensations to the locals.

Upon encountering an arising issue, the majority of inexperienced researchers will crumble while trying to find a solution. Through years of accruing real life experience in raw data investigation, we have devised a strategy of “*Di bat bien ung van bien*” (using unchanged principles to respond to and or manage the change), which involves 8 critical points: ***Safety, Quality, Excitement, Rapport, Efficiency, Note taking, Efficacy, Proactiveness***

- **Safety** of the survey crew is of utmost importance, to avoid any potential risks to each member
- For questionnaires, **quality** is prioritized over quantity. Any form collected must be of high standard
- Strive to maintain a high level of enthusiasm and **excitement**, as well as harmony within the crew
- Build and maintain good **rapport** with local authorities and citizens

- Achieve **cost efficiency** on work related matters
- **Keep notes** of daily activities and information of each crew member, as well as an analysis and recap of each day's workload
- Thorough in preparation, calm and flexible in execution, decisive in action. Use **efficacy** as a unit of measurement in working
- **Proactive** in all scenarios, always come prepared with back-up plans

In summary, success in the data collection phase for his research is heavily dependent on real life investigations. In spite of the importance of internal factors such as labor force, means of transportation and financial resources, scientists ought not to overlook external factors or arising matters that can impede the research's safety, schedule or quality. "*Di bat bien ung van bien*" is the key to overcome these problems.

Review questions

1. Present procedures of final survey in detail
2. Present principles to overcome difficulties and or challenges encountered over the course of final survey

Chapter 7. Data entry and coding

Preprint DOI: 10.31219/osf.io/kxp6b

Date: February 3, 2021

Abstract

Inputting and coding data are the last step in the primary data investigation ecosystem. This task is often time-consuming given the demand for accuracy, completeness and easy-to-use of the information in the data sheet. In order to help readers, especially young scientists with little or no experience in primary data investigation, to access and practice the aforementioned step effectively, in this chapter, we will present systematically the process involved in data entry and coding. Specifically, this includes designing an input table (codebook and data sheet) and checking the data.

Keywords: primary data, input table, codebook, variables

Introduction

After an arduous and costly procedure from questionnaire design to pilot and formal survey [7,16,23], researchers now have a set of completed questionnaires. However, the obtained information is discrete, "dead" since a single piece can add little value to the research. In general, a typical survey receives a large number of responses, ranging from hundreds to thousands [5,21]. They are then synthesized to yield primary data that will be processed using calculations, statistical models, thereby producing scientific information and data. To achieve that dataset, collected answers must undergo a process of input and coding, adhering to certain rules and

standards so that users can understand the "literal", encoded information, and vice versa. Here are 3 steps in inputting and coding data:

- Step 1. Design an input table (codebook and data sheet)
- Step 2. Input data (data entry table)
- Step 3. Check the data (check for accuracy, completeness and comprehensibility).

Designing an input table

Designing an input table is the first step, involving (1) building the codebook and (2) building the data sheet. Although the focus of this task is on the codebook and data sheet, rechecking the questionnaire is indispensable. In other words, the questionnaire is the overarching point of all activities. It is through the questionnaire that researchers can classify information and build the codebook.

Building the codebook

Data encryption/digitization is understood as the information coding step. Each question will be encoded as a “variable” whose information corresponds with that of each question [18]. These variables will be included in the model (e.g., econometric model) for analysis. As a result, the scientist will clearly identify what type of variable each piece of information belongs to (continuous variable, ordinal variable, binary variable, or nominal variable). The codebook is responsible for explaining all the information, codes of questions, answers in the questionnaire and the input table [8,11]. The goal is to enable anyone, upon accessing the data sheet, to understand what it describes, the meaning and effects of the variables and so forth. A detailed description of the information field in a codebook is provided in Table 6.

Table 6. Information field codebook

#	Dimensions	Example 1	Example 2	Example 3
1	Question ID	A1	A2	A3
2	Question content	Which district are you living in Hanoi?	To what extent do you feel about the air quality in your area?	Does the air quality vary across districts in Hanoi?
3	Variable	Location	AirDistrict	AirVsOtherDistrict
4	Variable Characteristics	Binary	Numerical on five points 1-5 scale	Binary
5	Code	(1, 0)	(1, 2, 3, 4, 5)	(1, 0)
6	Meaning	Living district of the interviewees	(Very good, good, normal, bad, very bad)	(Yes, No)

Sources: Survey on environmental pollution in Hanoi [8]

Table 01 presents six information fields, including: question ID, content, variable, variable characteristics, code, and meaning. First, the ID. Each question has a unique ID to be distinguished from one another in the questionnaire. And each of them has content. For example, in question A1, the content is “Which district are you living in Hanoi?”. Next is the variable, representing symbols for each question. As mentioned above, each question corresponds to a variable in the econometric model. It is noted that variables should be succinct, enough for readers to understand the question. Usually

it is advisable not to include more than three acronyms of three keywords for each variable. An example is the variable “Location” in question A1. The characteristics of a variable is its type that is often nominal, or ordinal, or continuous, or binary. In example 1, "Location" is a binary variable. And the code value of this variable is either 1 or 0. The last information field explains the meaning of the code value. Back to example 1, the meaning indicates where the interviewee lives, with value 1 in inner-city areas and value 0 in the suburbs. Likewise, in question A2, the content is to find out people’s opinions on the air quality in their neighborhood. The variable is “AirDistrict”, which is categorized as ordinal, measuring from 1 to 5 corresponding to the air quality from very good to very bad.

Data sheet

A data sheet is designed to store all the useful information of collected answer sheets from the official survey. Building the input table is not as difficult and complicated as the questionnaire since it is based on the already existing content of the questionnaire. However, there is a set of requirements and standards that needs to be met: complete, clear, scientific. First, it should contain all the useful information of collected answers. Clarity means it is easy to code and does not cause confusion in the input process. The scientific characteristic is shown in the logical arrangement of rows, columns, and groups of information, ensuring ease of reading, synthesis and calculation. An input table is a matrix of $m \times n$ (m is the number of questions, n the number of sheets, or observations). The two top rows present the questions’ number, the interviewees' ID and the variable name (Figure 6). Information from collected answers will be filled in accordingly and in the correct box.

	A	B	C	D	E	F	G	H
1			A1	A2	A3	A4	A5	A6
2	No.	Interview ID	Location	AirDistrict	AirVsOtherDistrict	AirCity	AirCityVsClosedCity	AirCityVsSouthCity
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								

Figure 6. An unfilled-in data sheet

Data entry

After designing the input table and the codebook, the researcher instructs the input participants to understand and perform the task correctly (Figure 7). Normally, the survey participants will input the information in the questionnaire that they have investigated, or exchange sheets with each other so as to avoid subjective errors. This is quite convenient because after a period of hands-on experience, they will grasp the questions as well as the information in the questionnaire, hence assisting data entry efficiently.

	A	B	C	D	E	F	G	H
1			A1	A2	A3	A4	A5	A6
2	No.	Interview ID	Location	AirDistrict	AirVsOtherDistrict	AirCity	AirCityVsClosedCity	AirCityVsSouthCity
3	1	TH01	Tay Ho	5	1	4	3	3
4	2	TH02	Tay Ho	2	1	4	4	4
5	3	TH03	Tay Ho	3	1	4	4	4
6	4	TH04	Tay Ho	2	1	4	3	4
7	5	TH05	Tay Ho	3	1	4	3	3
8	6	TH06	Tay Ho	2	1	4	4	4
9	7	TH07	Tay Ho	4	1	3	3	1
10	8	TH08	Tay Ho	2	1	3	4	4
11	9	TH09	Tay Ho	3	1	4	4	4
12	10	TH10	Tay Ho	3	1	4	3	3

Figure 7. A filled-in data sheet [8]

Data check

Filled-in data requires checking to avoid errors or missing information. In fact, anyone can make mistakes when entering data as there exists minor errors that are hard to detect. Therefore, it takes time to meticulously check each and every information field for each and every observation. Although the scientist has “cleaned up” the data, in this step, should there be any obvious errors, they need to be checked and eliminated. Specifically, the data is checked according to each variable. Easily detectable errors include missing and wrong information, which might create a value out of the normal range or that is too large (like a value of 6 in the 1-to-5 value range, an error easy to spot); typo, not according to the codebook (like “w3” instead of “3” due to the close space between “w” and “3” in the keyboard). Any abnormal points must be checked and compared with the answer sheet, all the entered data in accordance with the value presented in the sheet as well. It is noteworthy that the data might be unusual or unreasonable, yet it will be processed in the cleaning and analysis step.

Some notes on data entry and checking

- First, the most important criterion is accurate, complete and easy to understand.
- Second, this step should be done as soon as possible, preferably within the "first week" of the official survey. The longer it is left undone, the poorer the quality of the entry.
- Third, the detection of errors (missing, wrong) should be done coincidentally with the data entry process. There is no need to wait until the end of the entry process to check the data.
- Fourth, a person simply cannot check all errors perfectly. Hence, in order to minimize errors in the data entry process, the leader should divide the team into two data entry groups. Then the two groups will work together, compare the results, find out the errors and agree on the filling methods.

In summary, data entry and coding are the last and very important step in the primary data investigation “ecosystem”. This step helps researchers to obtain a set of interconnected datasets, ready for later statistical analysis. Collected information is encoded, according to certain rules and standards, from “word” to “keywords” corresponding to each variable type. The first task is building a clear, detailed and scientific input table. The variables are encoded based on the codebook to serve statistical analysis and process. After data entry, there are two common problems: wrong and missing information. Thus, under specific circumstances, depending on the characteristics of the data, researchers need to have effective and scientific process methods.

Review questions

1. List and present steps in the data entry process.
2. Present the content of the data checking step.

3. Present some notes on data designing, data entry and data checking.

References

- [1] V. Rao, Data is king. The one who masters Data with Science will rule the World! So are You Ready for Data Science?, Linkedin. (2018). <https://www.linkedin.com/pulse/data-king-one-who-masters-rule-world-so-you-ready-vinay-rao-mle-/> (accessed February 2, 2021).
- [2] Q.H. Vuong, V.P. La, T.T. Vuong, M.T. Ho, H.K.T. Nguyen, V.H. Nguyen, H.H. Pham, M.T. Ho, Data descriptor: An open database of productivity in Vietnam's social sciences and humanities for public use, *Sci. Data*. 5 (2018) 1–15. <https://doi.org/10.1038/sdata.2018.188>.
- [3] Editorial Team, The Exponential Growth of Data, Insidebigdata. (2017) <https://www.un.org/en/sections/issues-depth/big-da>. <https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/#:~:text=Human- and machine-generated data,growth toward 2020 and beyond>. (accessed February 2, 2021).
- [4] V.C. Dam, *Phuong phap luan nghien cuu khoa hoc (Scientific research methodology)*, NXB Khoa hoc ky thuat, Hanoi, Vietnam, 1999.
- [5] R.B. Uma Sekaran, *Research Methods For Business: A Skill Building Approach*, 7th Edition, 2016. https://doi.org/10.1007/978-94-007-0753-5_102084.
- [6] J.J. Hox, H.R. Boeije, Data Collection, Primary vs. Secondary, *Encycl. Soc. Meas.* (2004) 593–599. <https://doi.org/10.1016/B0-12-369398-5/00041-4>.
- [7] Kenneth D. Bailey, *Methods of Social Research*, fourth ed., The Free Press, New York, USA., 1994.
- [8] Q. Van Khuc, T.V. Phu, P. Luu, Dataset on the Hanoian suburbanites' perception and mitigation strategies towards air pollution, *Data Br.* 33 (2020) 106414. <https://doi.org/10.1016/j.dib.2020.106414>.
- [9] Q. Van Khuc, T.A.T. Le, T.H. Nguyen, D. Nong, B.Q. Tran, P. Meyfroidt, T. Tran, P.B. Duong, T.T. Nguyen, T. Tran, L. Pham, S. Leu, N.T. Phuong Thao, N. Huu-Dung, T.K. Dao, N. Van Hong, B.T. Minh Nguyet, H.S. Nguyen, M.W. Paschke, Forest cover change, households' livelihoods, trade-offs, and constraints associated with plantation forests in poor upland-rural landscapes: Evidence from north central Vietnam, *Forests*. 11 (2020). <https://doi.org/10.3390/F11050548>.

- [10] Q.H. Vuong, T.K. Nguyen, Data on Vietnamese patients' financial burdens and risk of destitution, *Data Br.* 9 (2016) 543–548. <https://doi.org/10.1016/j.dib.2016.09.040>.
- [11] P.A. Champ, K.J. Boyle, T.C. Brown, *A Primer on Nonmarket Valuation*, 2017. <http://link.springer.com/10.1007/978-94-007-7104-8>.
- [12] V.T. Nguyen, *Tu nghien cuu den cong bo* (from research to publication), NXB tong hop tp Ho Chi Minh, Ho Chi Minh city, Vietnam, 2013.
- [13] Q. Van Khuc, M. Alhassan, J.B. Loomis, T.D. Tran, M.W. Paschke, Estimating Urban Households' Willingness-to-Pay for Upland Forest Restoration in Vietnam, *Open J. For.* 06 (2016) 191–198. <https://doi.org/10.4236/ojf.2016.63016>.
- [14] Q.H. Vuong, Data Descriptor: Survey data on Vietnamese propensity to attend periodic general health examinations, *Sci. Data.* 4 (2017) 1–10. <https://doi.org/10.1038/sdata.2017.142>.
- [15] Bill Gillham, *Developing a Questionnaire*, CONTINUUM, London, UK, 2000.
- [16] T.D. Tran, Q. Van Khuc, Primary data, OSF Prepr. (2021) 1–6. <https://doi.org/10.31219/osf.io/f25v7>.
- [17] Q.-H. Vuong, P. Pham, M.-H. Nguyen, C.-T. Ngo, P.-M. Tran, Q. Van Khuc, Farmers' livelihood strategies and perceived constraints from poor and non-poor households: A dataset from a field survey in Nghe An, Vietnam Quan-Hoang Vuong, *Data Br.* (under Rev. (2021) 1–18.
- [18] T.D. Tran, P. Pham, Q. Van Khuc, Questionnaire design, OSF Prepr. (2021) 7–13. <https://doi.org/10.31219/osf.io/q3um6>.
- [19] M. Bloor, J. Frankland, M. Thomas, K. Robson, *Focus Groups in Social Research*, *Focus Groups Soc. Res.* (2001). <https://doi.org/10.4135/9781849209175>.
- [20] C. Work, *danielmediamanchestercollege*, (2012). <https://danielmediamanchestercollege.wordpress.com/2012/10/22/audience-feedback/>.
- [21] T. Tran, L. Pham, Q. Khuc, Sampling, OSF Prepr. (2021) 21–28. <https://doi.org/10.31219/osf.io/w9qks>.
- [22] P. Pham, T. Tran, Q. Khuc, Focus group, OSF Prepr. (2021) 14–20. <https://doi.org/10.31219/osf.io/nfczd>.
- [23] P. Tran, T. Nguyen, T. Tran, Q. Khuc, Pilot survey, OSF Prepr. (2021) 29–33. <https://doi.org/10.31219/osf.io/dwhja>.