# 1

# Data Collection in Education



## 1.1. Use of existing databases in education

A piece of data is an elementary description of a reality. It might be, for instance, an observation or measurement. A data point is not subject to any reasoning, supposition, observation or probability, even though the reason for and method of its collection are non-neutral. Indisputable or undisputed, it serves as the basis for a search, any examination, expressed in terms of a problem. Data analysis is normally the result of prior work on the raw data, imbuing them with meaning in relation to a particular problem, and thereby obtaining information. Data are a set of measurable values or qualitative criteria, taken in relation to a reference standard or to epistemological positions identified in an analysis grid. The reference grid used and the way in which the raw data are processed are explicit or implicit interpretations, which can transform (or skew) the final interpretation. For example, discretizing data (classifying them by establishing boundaries) in a graph enables an analyst to associate a meaning (an interpretation) with those data, and thus create new information in relation to a given problem.

Societies are restructuring to become "knowledge societies" in a context where in the global economy, innovations are based on the storage and exploitation of knowledge (value creation), and on the training and qualifications of actors in knowledge exploitation. Today, our knowledge society is faced with an explosion in the volume of data. The volume of the data created and copied is increasing exponentially, and the actors in civil society and in economics are taking great pains to find solutions allowing them to manage, store and secure those data. The tendency toward the use of *Big Data* to obtain a competitive edge and help organizations to achieve their objectives requires the collection of new types of information (comments posted on websites, pharmaceutical testing data, researchers' results, to cite only a few examples) and examination of the data from every possible

angle in order to reach an understanding and find solutions. Education is no exception to this rule, and burgeoning quantities of data are available in this field. Though this is not an exhaustive list, we could point to international databases with which practitioners, students, researchers and businesses can work. Data enables us to act on knowledge, which is a determining factor in the exercise of power. It should also be noted that in today's world, data transmission is subject to rigorous frameworks (for example, issues of data privacy).

Below, we give a few examples of the databases that can be used.

### 1.1.1. *International databases*

The available international databases are, primarily, those maintained by intergovernmental organizations. For example, the UNESCO Institute for Statistics and the World Bank run a paying service for statistical monitoring in education. They facilitate international comparisons. The main access points are listed in Table 1.1.

| | |
|---|---|
| UNESCO database: education | http://www.uis.unesco.org/education |
| World Bank database: education | http://datatopics.worldbank.org/education/ |
| OECD database: education | http://www.oecd.org/education/ |
| United Nations database: for teachers | https://www.gapminder.org/for-teachers/ |
| Eurostat database: education and training for Europe-wide comparisons and longitudinal monitoring | http://ec.europa.eu/eurostat/web/education-and-training/data/database |
| Eurydice database: institutional information network on educational policy and systems in Europe | http://www.eurydice.org/ |
| Canadian databases on education the world over | https://www.wes.org/ca/wedb/ |
| Database of National Center for Education Statistics | https://en.wikipedia.org/wiki/National_Center_for_Education_Statistics |

**Table 1.1.** *International databases*

| | |
|---|---|
| Educsol database: education | http://eduscol.education.fr/numerique/dossier/archives/ressources-en-ligne/bases/bases-donnees |
| INSEE database: teaching and education by commune | http://www.insee.fr/fr/themes/theme.asp?theme=7 |
| Databases of the Ministry for National Education and the Ministry for Higher Education and Research | http://www.education.gouv.fr/bcp/mainFrame.jsp?p=1 https://data.enseignementsup-recherche.gouv.fr/explore/?sort=modified |
| Statistical data from CAPES (French teacher training qualification) examination | http://www.devenirenseignant.gouv.fr/cid98479/donnees-statistiques-des-concours-du-capes-de-la-session-2015.html |

**Table 1.2.** *French national databases*

### 1.1.2. *Compound databases*

The list offered here is by no means exhaustive, and the situation is constantly changing. It shows the main databases compiled by organizations, communes and researchers.

| | |
|---|---|
| Institutional databases: e.g. that of the Académie d'Aix-Marseille | https://fichetab.ac-aix-marseille.fr/ |
| Emmanuelle database: compiles all editions of textbooks published in France since 1789, for all disciplines and all levels of education | http://www.inrp.fr/emma/web/index.php |

**Table 1.3.** *Compound databases*

Researchers establish numerous databases. A few examples are presented in Table 1.4.

| | |
|---|---|
| Database of the Center for Study and Research on qualifications | http://mimosa.cereq.fr/reflet/index.php?lien=educ_pres |
| Databases of rural schools of the Observatory for Education and Territory, longitudinal examination | http://observatoireecoleetterritoire |

**Table 1.4.** *Databases established by researchers*

An analyst working with the data can then construct their own statistical tables and apply their own processes to obtain an answer to a given question.

## 1.2. Survey questionnaire

In the field of education, it is very common to use questionnaires to gather data regarding educational situations. Think, for example, about the areas of guidance counselling, for school and for work; about the paths taken by the pupils; the social make-up of the community of teachers and parents; knowledge about the establishments; etc.

In all cases, in order to write a useful questionnaire, it is important to know the general rules. Below, we very briefly discuss the general principles, but there are numerous publications in sociology of investigation, which must be consulted to gain a deeper understanding.

### 1.2.1. *Objective of a questionnaire*

To begin with, it is necessary to very clearly specify the aim of a questionnaire, and to set out the requirements in terms of collecting information. Indeed, questionnaires are always used in response to a desire to measure a situation – in this instance, an educational situation – and their use is part of a study approach which is descriptive or explicative in purpose, and quantitative in nature.

The aim, then, is to describe a population or a target subpopulation in terms of a certain number of criteria: the parents' socioprofessional profile, the pupils' behavior, the intentions of the educational actors, the teachers' opinions, etc. The goal is to estimate an absolute or relative value, and test relations between variables to verify and validate hypotheses.

In any case, the end product is a report, whatever the medium used. The report is made up of a set of information having been collected, analyzed and represented.

As with any production, the conducting of an inquiry must follow a coherent and logical approach to a problem.

### 1.2.2. *Constitution of the sample*

Very simply, there are two main ways of constructing a sample for an investigation. Either the entire population is questioned – e.g. all the pupils in a school – in which

case, the study is said to be exhaustive or else only a portion of the population is questioned – e.g. just a handful of pupils per class – in which case, the study is done by survey.

However, the technique of survey-based investigation necessitates a degree of reflection on the criteria used to select the portion of the population to be surveyed. That portion is called the sample. In order to obtain reliable results, the characteristics of the sample must be the same as those of the entire population. There are two sampling methods. To begin with, there are probabilistic methods, meaning that the survey units are selected at random. These methods respect statistical laws. With simple random surveying, a list of all survey units is drawn up, and a number are selected at random. With systematic sampling, from the list of all survey units, we select one in every $n$ number. With stratified sampling, if the population can be divided into homogeneous groups, samples are randomly selected from within the different groups.

Second, we have non-probabilistic methods. The best-known non-probabilistic method is the quota method, which is briefly described below.

This method involves four steps: studying the characteristics of the base population in relation to certain criteria of representativeness, deducing the respective role of these various criteria in relative value, determining a sampling rate to determine the sample size and applying the relative values obtained to the sample.

The example below illustrates the principle: consider a population of 10,000 residents. The analysis of that population (INSEE) shows that there is 55% of women, 45% of men, 10% aged under 20, 20% aged between 20 and 40, 25% aged between 40 and 60, and 45% over 60.

These percentages are referred to as quotas. If a sampling rate of 1/20 is chosen to begin with, this means that the ratio “sample size/size of population under study” must be 1/20.

The sample size, then, is 10,000/20 = 500 people. The structure of the sample is determined by applying the quotas. There will be 500 × 55% = 275 women, 500 × 45% = 225 men, 500 × 10% = 50 under the age of 20, 500 × 20% = 100 between the ages of 20 and 40, 500 × 25% = 125 people between 40 and 60 years of age, and 500 × 45% = 225 people over 60.

The difficulty lies in setting the sampling rate. An empirical method is to estimate that the sampling rate must be such that the smallest group obtained is at

least 30 people. For probabilistic methods, there are formulae in place which relate to statistical functions.

Probabilistic methods allow us to determine the sampling rate as a function of the population size with a sufficiently large confidence interval.

NOTE.– We can also calculate the overall sample size using statistical concepts, which we shall touch upon later on in this book. Therefore, the presentation here is highly simplified.

The sample size that should be chosen depends on the degree of accuracy we want to obtain and the level of risk of error that we are willing to accept.

The accuracy is characterized by the margin of error "e" that is acceptable. In other words, we are satisfied if the value of the sought value A is between A (1 – e) and A (1 + e). Often, e is taken to be equal to 0.05.

The risk of mistake that we are willing to accept is characterized by a term "t", which will be explained in Chapter 3. If we accept conventional 5% risk of error, then t = 1.96.

There are then two approaches to quickly calculate the size of a sample:

– based on a proportion, we can calculate the sample size using the formula:

$$n = \frac{t^2 xp(1-p)}{e^2}$$

where:

– n = expected sample size;

– t = level of confidence deduced from the confidence rate (traditionally 1.96 for a confidence rate of 95%) – a reduced centered normal law;

– p = estimated proportion of the population exhibiting the characteristic that is under study. When that proportion is not known, a preliminary study can be carried out, or else the value is taken to be p = 0.5. This value maximizes the sample size;

– e = margin of error (often set at 0.05).

Thus, if we choose p = 0.5, for example, choosing a confidence level of 95% and a margin of error of 5% (0.05), the sample size must be:

$$n = 1.96^2 \times 0.5 \times 0.5 / 0.05^2 = 384.16 \text{ individuals}$$

It is then useful to verify whether the number of individuals per quota is sufficient based on a mean. In order to do so, we need an initial estimation of the standard deviation, so we can adjust the sample on the basis of the accuracy of results it obtains and the expected level of analysis:

$$n = \frac{t^2 x \sigma^2}{e^2}$$

where:

– n = expected sample size;

– t = level of confidence deduced from the confidence rate (traditionally 1.96 for a confidence rate of 95%) – a reduced centered normal law;

– $\sigma$ = estimated standard deviation of the criterion under study. This value will be defined in Chapter 2;

– e = margin of error (often set at 0.05).

### 1.2.3. *Questions*

#### 1.2.3.1. *Number of questions*

Apart from exceptional cases, if the surveys are being put to people on the street, the questionnaire must be reasonably short: 15-odd questions at most.

If the respondents are filling in a questionnaire in an educational institution or at home, the number of questions may be greater.

#### 1.2.3.2. *Order of the questions*

A questionnaire should be structured by topic, and presented in the form of a progression, leading from the fairly general to the very specific. Personal questions (age, address, class, gender, etc.) must be posed at the end of the questionnaire.

#### 1.2.3.3. *Types of questions*

Remember the different types of questions:

– closed-ended, single response;

– closed-ended and ranged: e.g. enter a value between “1 and n” or between “very poor and very good”.

NOTE.– These questions must offer an even number of choices when the goal is to express an opinion; otherwise, there is a danger that the responses will be concentrated around the central option:

– closed-ended, multiple choice: one or more responses out of those offered;

– closed-ended, ordered: the responses are ranked in order of preference;

– open-ended: the respondent has complete freedom to respond in text or numerical form.

It is worth limiting the number of open-ended questions, because they take a long time to unpick with a survey software tool. As far as possible, it is preferable to reduce an open-ended question to a multiple-choice one.

### 1.2.4. *Structure of a questionnaire*

The questionnaire must contain three parts: the introduction, the body of the questionnaire and the conclusion.

#### 1.2.4.1. *Introduction and conclusion*

The questionnaire needs to be formulated in a way that will draw the respondent in. It generally includes greetings, an introduction to the researcher, the context and the purpose of the study, and an invitation to answer the questions that follow.

EXAMPLE.– (Greeting), my name is Y; I am an intern at company G. I am carrying out a study on customer satisfaction. Your answers will be very helpful to us in improving our service.

The conclusion is dedicated to giving thanks and taking leave. Once written, the questionnaire will be tested in order to reveal any difficulties in understanding the questions or difficulties in the answers.

#### 1.2.4.2. *Body of the questionnaire*

The questions must be organized logically into several distinct parts, from the most general questions to the most specific ones. Generally, information questions or personal identification ones (e.g. age, gender, socioprofessional status, class, etc.) are posed at the end.

### 1.2.5. *Writing*

#### 1.2.5.1. *General*

The writing of the questionnaire in itself requires great rigor. The vocabulary used must be suited to the people who are being questioned. It is important to use simple words in everyday language and avoid overly technical or abstract words or ambiguous subjects.

The physical presentation must be impeccable and comfortable to read. The questions should be aligned with one another and, likewise, the grids used for the answers. Word-processing tools generally perform very well in this regard.

#### 1.2.5.2. *Pitfalls to be avoided*

The questions posed must invite non-biased responses. Questions that might bring about biased responses are those that involve desires, memory, prestige and the social environment. Generally speaking, the questions should not elicit the response ("Don't you think that…") or contain technical, complicated or ambiguous terms. They must be precise. Indeed, any general question will obtain a general response that, ultimately, gives us very little information.

Below, we present a series of errors that are to be avoided.

The question "Do you play rugby a lot?" is too vague and too general. It could usefully be replaced by "How many times a week do you play rugby?".

"What is your favorite activity?" This question strips the respondent of the freedom not to have a favorite activity. "Do you have a favorite activity?" is better, because it leaves the respondent with the freedom to have one or not.

The question "Do you have daughters, and what age is your youngest?" contains a twofold question. It needs to be broken up into two separate questions.

The question "Don't you agree that adults do not do enough walking?" lacks clarity, because of the double negative and the length of the question. The respondent may not actually have an opinion on the issue, or might respond with a "yes" or "no" carrying little meaning. It would be useful to replace this question with one inviting a scaled response.

"Do you eat vegetables often?" is a bad question, because the word "often" may be interpreted differently by the various people to whom it is put. It is preferable to use a scaled question:

□ every day □ twice a week □ once a week □ etc.

## 1.3. Experimental approaches

### 1.3.1. *Inductive and deductive approaches*

In sections 1.1 and 1.2, we have shown how to obtain data, either by directly consulting the databanks available in the pre-existing literature or by compiling original data based on investigations. In both cases, the researcher acts as an impartial observer, doing nothing more than collecting the data. Theoretically, they should have no direct influence on what they are observing. Once the data have been harvested, the researchers analyze them and attempts, *a posteriori*, to declare results to establish patterns, rules and laws. The approach is inductive. If we observe one variable in particular, and attempt to establish a link with another variable, the link we see is invariably fogged by the presence of parasitic variables.

Experimentation is another way to obtain data. In this case, the approach is deductive, and the researcher is no longer a mere observer. They begin by formulating an initial hypothesis, called the research hypothesis, and implement a procedure and an artificial device to validate or undermine the original hypothesis. They attempt to control the situation in order to see how a variable, known as the dependent variable, responds to the action of one or more other variables, known as independent variables, which the researcher can manipulate, thereby circumventing the problem of parasitic variables, as far as possible. Yet in the human sciences, unlike in the hard sciences, this control is never absolute.

Often, an experiment is cumbersome to put in place and complex to manage; the impact of this is to limit the size of the samples observed. The small size of these samples constantly forces us to wonder: "Had I worked on a different sample, would I still have obtained the same result?" In order to answer this question, we have to use confirmatory statistics (see Chapter 3), the aim of which is to determine whether the results obtained from the sample can be generalized to the whole population.

Below, we present an example drawn from cognitive psychology, which is a pioneering discipline in the field and frequently uses the experimental approach, before going on to talk about the form that the experimental approach might take for research in education.

### 1.3.2. *Experimentation by psychologists*

Psychologists are adept with the experimental approach – particularly in cognitive psychology. They place the individuals in their sample in situations that favor the observation of a link between a dependent variable and an independent one. To illustrate this approach, below, we present an example drawn from Abdi [ABD 87].

The study relates to the effect of the use of mental imagery for memorization. The question posed is: do subjects have better retention of information memorized with the help of images than information memorized without images? To answer this question, an experiment is set up. Subjects are asked to learn pairs of words, such as "bicycle-carrot" or "fish-chair". They then need to give the associated second word – for instance, if the tester says "carrot", they should answer "bicycle", and so on.

Two groups of 15 subjects are set up: the experimental group (EG) and the control group (CG). Each of the groups is shown the pairs of words on a screen, represented by an image for the EG and the written words for the CG. Afterwards, they are questioned about what they have memorized.

The dependent variable is the "number of word pairs memorized", and the independent variable is the "learning method", with two facets: "with images" and "without images".

In this study, the experiment demonstrated that memorization was far greater in the EG than in the CG. Thus, it appears that the more visual method aids memorization.

### 1.3.3. *Experimentation in education*

As we shall show later on, research in education can take multiple and diverse paths, but all sharing a common goal: to collect data in order to draw conclusions about certain phenomena pertaining to the human sciences in general, and education in particular. One of these avenues is experimentation, and the aim is to construct theoretical models that can help understand, and even predict, different aspects of education, and particularly to improve educational practices. Measurement of the "effectiveness of teaching" is one of the fundaments of such work.

Below, to demonstrate the form an experiment may take, we present two situations in the field of research in education.

#### 1.3.3.1. *A simple example: fictional, but realistic*

We wish to test the effectiveness of a new method for teaching reading, and to find out whether it is more effective than the method conventionally used.

##### We construct an experiment

Consider two primary school classes with 26 and 23 pupils, respectively. The first, the CG, will learn with the conventional method, while the second, the EG, learns with the new method.

After 2 months of teaching, an identical test is put to both classes. An assessment grid is used to precisely grade each pupil, and that grade serves as a performance indicator. We may decide that the performance indicator for each of the two classes is the mean obtained by the pupils. In this case, the dependent variable is the "mean grade obtained" and the independent variable is the "group", which has two facets: CG and EG.

By comparing the two means, we can tell which of the two groups has performed better. However, is it the learning method which has made the difference, or other, parasitic variables?

Before drawing a radical conclusion, there are a number of factors that need to be taken into account such as the initial "level" of the two classes, the teacher's influence (teacher effect), the sample size and, indubitably, other parasitic variables as well. Although we cannot absolutely avoid parasitic effects, it is wise to identify them and minimize their effects as far as possible. For example, to ensure that the two classes are of the same level to begin with, we could choose them in the wake of preliminary testing before the start of the experiment. For the teacher effect, while it is not always possible, the ideal is that the same teacher should dispense the education to both classes.

#### 1.3.3.2. *A real-world example in the context of an experiment at university*

The real-world example on which we wish to focus relates to a third-year undergraduate course in engineering sciences, in the sector of mechanical engineering. It compares the effectiveness of conventional teaching techniques (classes and supervised work) with the same education, on the same content, but using problem-based learning (PBL). In PBL, students work in small groups. They are given a problem to solve and must pull together the information necessary to solve it, and even independently acquire the essential new concepts, whereas in conventional education, new knowledge is imparted in class and through guided work.

The questions which are posed are as follows:

– is the acquisition of fundamental academic knowledge as effective, in PBL, as it is in conventional education?

– is the memorization of knowledge as significant, in PBL, as in conventional education?

In other words, we want to know whether it is more effective to obtain fundamental academic knowledge independently (through PBL) or by structured teaching (classes and guided work), and also find which of the two approaches best aids memorization.

### We construct an experiment

We have two groups: the EG (54 students), working by PBL; and the CG (46 students), following the usual path of classes and directed work. To answer the first question, a longitudinal measure is put in place, running from the start of March to the end of May. It mainly comprises a general level test before the course starts (pretest) and a test after the course (posttest) relating to the discipline-specific academic content retained from the course.

In the pretest, the two groups obtain an equivalent average score, and the posttest shows that after the course, the two groups are still equivalent, on average. However, a more detailed analysis of the results clearly shows that the best students in the EG have made more progress than the best students in the CG, and conversely, that the weakest in the EG have less success than the weakest in the CG. This experiment shows that when it comes to fundamental academic teaching, problem-based learning is more effective for the stronger students, but risks leaving the weaker ones behind.

To answer the second question, a memorization test is organized for when the students return in September. The results demonstrate that, overall, the EG was better able to memorize the knowledge than the CG.

In all cases, data are recorded about the individuals, and to begin with, those data are entered into a binary table (Individual/Variable). The same is true for all data collection methods.

This table is not legible. Hence, the rest of this book examines the various approaches that can be used to extract information from it, formulate hypotheses and even draw conclusions.