Quantitative Methods and Statistics

true

Version compiled 22 Oct 2020

Contents

Preface

Data are becoming ever more important, in all parts of society, including academia, and including the humanities. The availability of large amounts of digital data (such as text, speech, video, behavioural measurements) raises new research questions, which are typically and often investigated using quantitative methods. Aimed at humanities researchers and students, this book offers an overview of and introduction into the most important quantitative methods and statistical techniques used in the humanities. The book provides a solid methodological foundation for quantitative research, and it introduces the most commonly used statistical techniques to describe data and to test hypotheses. This will also enable the reader to critically evaluate such quantitative research.

This textbook is being used in the course *Methods and Statistics 1* at Utrecht University (Linguistics program). The book is also highly suitable for self-study at a basic level, for everybody who wishes to learn more about quantitative methods and statistics.

The main text has been kept free of mathematical derivations and formulas, which are typically not very helpful for humanities scholars and students. Our explanation is rather conceptual, and rich in examples. Where necessary we present derivations and formulas in separate sections.

This book also contains instructions on how to "do" the statistical analyses and visualisations, both in SPSS (version 22 or later) and in R (version 3.0 or later). These instructions too are in separate sections.

We would like to thank our co-teachers in various courses for the many discussions and examples that have been used in any shape or form in this textbook. We thank our students for their curiosity and for their sharp eyes in spotting errors and inconsistencies in previous versions.

We are also thankful to Gerrit Bloothooft, Margot van den Berg, Willemijn Heeren, Caspar van Lissa, Els Rose, Tobias Quené, Kirsten Schutter and Marijn Struiksma, for their advice, data, comments and suggestions.

We thank Aleksei Nazarov and Joanna Wall for translating this book from Dutch to English.

Utrecht, October 2020

Hugo Quené, https://www.hugoquene.nl

Huub van den Bergh, https://www.uu.nl/staff/HHvandenBergh

Notation

Following international usage we use the full stop (decimal point) as decimal separator; hence we write $\frac{3}{2} = 1.5$. Note that the decimal separator may vary between computers and between software packages on the same computer. Check which decimal separator is used by (each software package on) your computer.

License

Citation

Please cite this work as follows (in APA style):

Quené, H. & Van den Bergh, H. (2020). *Quantitative Methods and Statistics*. Retrieved 21 Oct 2020 from https://hugoquene.github.io/QMS-EN/ .

Technical details

All materials for this textbook are available at https://github.com/hugoquene/QMS-EN: this includes other versions of this textbook (EPUB, PDF, HTML), the source code (Rmarkdown and R) of the text including figures and examples, accompanying datasets used in the text, and figures as separate files.

The original Dutch version of this text was written in LaTeX, and was then converted to Rmarkdown, using `pandoc` (MacFarlane, 2020) and the `bookdown` (Xie, 2020) in Rstudio. The Dutch version is available at https://hugoquene.github.io/KMS-NL. The English translation is based on the Dutch LaTeX version (for Part I) and Rmarkdown version (for Parts II and III).

About the authors

Both authors work at the Faculty of Humanities at Utrecht University, the Netherlands. HQ is professor in the Quantitative Methods of Empirical Research in the Humanities, and he is also founding director of the Centre for Digital Humanities at Utrecht University. HvdB is professor in the Pedagogy and Testing of Language Proficiency, and he is also section chair in Dutch Language and Literature at the Dutch National Board of Tests and Examinations (CvTE).

Part I: Methodology

Chapter 1

Introduction

In this textbook, we will discuss the fundamental concepts, methods, and analytic techniques used in empirical scientific inquiry, both in general and as applied to the broad domain of language and communication. We will look at questions such as: What is a good research question? Which methodology is best for answering a given research question? How can researchers draw meaningful and valid conclusions from (statistical analyses of) their data? In this textbook, we will restrict ourselves to the most important fundamental concepts, and to the most important research methodologies and analytical techniques. In this first chapter, we will provide an overview of various types and forms of scientific research. In the following chapters, we will focus most of our attention on scientific research methodologies in which empirical observations are expressed in terms of numbers (quantitative), which may be analysed using statistical techniques.

1.1 Scientific research

To begin, we have to ask a question that refers back to the very first sentence above: what exactly is scientific research? What is the difference between scientific and non-scientific research (e.g., by investigative journalists)? Research conducted by a scholar does not necessarily have to be scientific research. Nor is research by journalists non-scientific by definition just because it is conducted by a journalist. In this textbook, we will follow this definition (Kerlinger and Lee, 2000, p.14):

> "Scientific research is systematic, controlled, empirical, amoral, public, and critical investigation of natural phenomena. It is guided by theory and hypotheses about the presumed relations among such phenomena."

Scientific research is systematic and controlled. Scientific research is designed such that its conclusions may be believed, because these conclusions are well-motivated. A research study can be repeated by others, which will (hopefully) lead to the same results. This demand that research be replicable also means that scientific research is designed and conducted in highly controlled ways (see Chapters **??** and **??**). The strongest form of control is found in a scientific experiment: we will therefore devote considerable attention to experimental research (§**??**). Any possible alternative explanations for the phenomenon studied are looked into one by one and excluded if possible, so that, in the end, we are left with one single explanation (Kerlinger and Lee, 2000). This explanation, then, forms our scientifically motivated conclusion on or theory of the phenomenon studied.

The definition above also states that scientific research is empirical. The conclusion a research draws about a phenomenon must ultimately be based on (systematic and controlled) observations of that phenomenon in reality – for example, on the observed content of a text or the behaviour observed in a test subject. If such observation is absent, then any conclusion drawn from such research cannot be logically connected to reality, which means that it has no scientific value. Confidential data from an unknown source or insights gained from a dream or in a mystical experience are not empirically motivated, and, hence, may not form the basis of a scientific theory.

1.1.1 Theory

The goal of all scientific research is to arrive at a theory of a part of reality. This theory can be seen as a coherent and consistent collection of "justified true beliefs" (Morton, 2003). These beliefs as well as the theory they form abstract away from the complex reality of natural phenomena to an abstract mental *construct*, which in its very nature is not directly observable. Examples of similar constructs include: reading ability, intelligence, activation level, intelligibility, active vocabulary size, shoe size, length of commute, introversion, etc.

When building a theory, a researcher not only defines various constructs, but also specifies the *relationships* between these constructs. It is only when the constructs have been defined and the relationships between these constructs have been specified that a researcher can arrive at a systematic explanation of the phenomenon studied. This explanation or theory can, in turn, form the basis of a *prediction* about the phenomenon studied: the number of spoken languages will decrease in the 21st century; texts without overt conjunctions will be more difficult to understand than texts with overt conjunctions; children with a bilingual upbringing will perform no worse at school than monolingual children.

Scientific research comes in many kinds and forms, which may be classified in various ways. In §**??**, we will discuss a classification based on paradigm: a

researcher's outlook on reality. Research can also be classified according to a continuum between 'purely theoretical' to 'applied'. A third way of classifying research is oriented towards the type of research, for instance, instrument validation (§1.3), descriptive research (§**??**), and experimental research (§**??**).

1.2 Paradigms

One criterion to distinguish different kinds of research is on the basis of the paradigm used: the researcher's outlook on reality. In this textbook, we have spent almost all of our attention on the empirical-analytical paradigm, because this paradigm has been written about the most and is the most influential. At present, this approach can be seen as 'the' standard approach, against the backdrop of which other paradigms try to distinguish themselves.

Within the *empirical-analytical* paradigm, we distinguish two variants: positivism and critical rationalism. Both schools of thought share the assumption that there exist lawful generalizations that can be 'discovered': phenomena may be described and explained in terms of abstractions (constructs). The difference between the two schools within the empirical-analytical tradition lies in the way generalizations are treated. Positivists claim that it is possible to make statements from factual observations towards a theory. Based on the observations made, we may generalize towards a general principle by means of induction. (All birds I have seen are also perceived by me to be singing, so all birds sing.)

The second school is critical rationalism. Those within this school of thought oppose the inductive statements mentioned above: even if I see masses of birds and they all sing, I still cannot say with certainty that the supposed general principle is true. But, say critical rationalists, we can indeed turn this on its head: we may try to show that the supposed general rule or hypothesis is not true. How would this work? From the general principle, we can derive predictions about specific observations by using deduction. (If all birds sing, then it must be true that all birds in my sample do sing.) If it is not the case that all birds in my sample sing, this means the general principle must be false. This is called the falsification principle, which we will discuss in more detail in **??**.

However, critical rationalism, too, has at least two drawbacks. The falsification principle allows us to use observations (empirical facts, research results) to make theoretical statements (regarding specific hypotheses). Strictly speaking, a supposed general principle should be immediately rejected after a single successful instance of falsification (one of the birds in my sample does not sing): if there is a mismatch between theory and observations, then, according to critical rationalists, the theory fails. But to arrive at an observation, a researcher has to make many choices (e.g., how do I draw an appropriate sample, what is a bird, how do I determine whether a bird sings?), which may cast doubt on the validity of the observations. This means that a theory/observation mismatch

could also indicate a problem with the observations themselves (hearing), or with the way the constructs in the theory (birds, singing) are operationalized.

A second drawback is that, in practice, there are very few theories that truly exclude some type of observation. When we observe discrepancies between a theory and observations made, the theory is adjusted such that the new observations still fit within the theory. In this way, theories are very rarely completely rejected.

One alternative paradigm is the critical approach. The *critical paradigm* is distinguished from other paradigms by its emphasis on the role of society; there is no one true reality: our image of reality is not a final one, and it is determined by social factors. Thus, insight into relationships within society, by itself, influences this reality. This means that our concept of science, as formulated in the definitions of research and theory given above, is rejected in the critical paradigm. Critical researchers claim that research processes cannot be seen as separate from the social context in which research is conducted. However, we must add that this latter viewpoint has lately been taken over by more and more researchers, including those that follow other paradigms.

1.3 Instrument validation

As stated above, research is a systematized and controlled way of collecting and interpreting empirical data. Researchers strive for insight into natural phenomena and into the way in which (constructs corresponding to) these phenomena are related to one another. One requirement for this is that the researcher be able to actually measure said phenomena, i.e., to express them in terms of an observation (preferable, in the form of a number). Instrument validation research is predominantly concerned with constructing instruments or methods to make phenomena, behaviour, ability, attitudes, etc. measurable. The development of good instruments for measurement is by no means an easy task: they truly have to be crafted by hand, and there are many pitfalls that have to be avoided. The process of making phenomena, behaviour, or constructs measurable is called *operationalization*. For instance, a specific reading test can be seen as an operationalization of the abstract construct of 'reading ability'.

It is useful to make a distinction between the abstract theoretical construct and the construct as it is used for measurements, which means: a distinction between the concept-as-intended and the concept-as-defined. Naturally, the desired situation is for the concept-as-defined (the test or questionnaire or observation) to maximally approach the concept-as-intended (the theoretical construct). If the theoretical construct is given a good approximation, we speak of an adequate or valid measurement.

When a concept-as-intended is operationalized, the amount of choices to be made is innumerable. For instance, the Dutch government institute that devel-

ops standardized tests for primary and secondary education, the CITO (Centraal instituut voor toetsontwikkeling, or Central Test Development Institute) must develop new reading comprehension tests each year to measure the reading ability exhibited by students taking the centralized final exams for secondary school students (eindexamens). For this purpose, the first step is to choose and possibly edit a text. This text cannot be too challenging for the target audience, but may also not be too easy. Furthermore, the topic of the text may not be too well-known – otherwise, some students' general background knowledge may interfere with the opinions and standpoints brought forward in the text. At the next step, questions must be developed in such a way that the various parts of the text are all covered. In addition, the questions must be constructed in such a way that the theoretical concept of 'reading ability' is adequately operationalized. Finally, exams administered in previous years must also be taken into consideration, because this year's exam may not differ too much from previous years' exams.

To sum up, a construct must be correctly operationalized in order to arrive at observations that are not only valid (a good approximation of the abstract construct, see Chapter **??**) but also reliable (observations must be more or less identical when measurement is repeated, see Chapter **??**). In each research study, the validity and reliability of any instance of measurement are crucial; because of this, we will spend two chapters on just these concepts. However, in instrument validation research, specifically, these concepts are absolutely essential, because this type of research itself is meant to yield valid and reliable instruments that are a good operationalization of the abstract construct-as-intended.

1.4 Descriptive research

Descriptive research refers to research predominantly geared towards describing a particular natural phenomenon in reality. This means that the researcher mostly aims for a description of the phenomenon: the current level of ability, the way in which a particular process or discussion proceeds, the way in which Dutch language classes in secondary education take shape, voters' political preferences immediately before an election, the correlation between the number of hours a student spent on individual study and the final mark they received, etc. In short, the potential topics of descriptive research are also be very diverse.

Example 1.1: Dingemanse et al. (2013) made or chose recordings of conversations in 10 languages. Within these conversations, they took words used by a listener to seek "open clarification": little words like *huh* (English), *hè* (Dutch), *ã?* (Siwu). They determined the sound shape and pitch contour of these words using acoustic measurements

> and phonetic transcriptions made by experts. One of the conclusions of this descriptive research is that these interjections in the various languages studied are much more alike (in terms of sound shape and pitch contour) than would be expected based on chance.

This example illustrates the fact that descriptive research does not stop when the data (sound shapes, pitch contours) have been described. Oftentimes, relationships between the data points gathered are also very interesting (see §1.1). For instance, in opinion polls that investigate voting behaviour in elections, a connection is often made between the voting behaviour polled, on the one side, and age, sex, and level of education, on the other side. In the same way, research in education makes a connection between the number of hours spent studying, on the one side, and performance in educational assessment, on the other side. This type of descriptive research, in which a correlation is found between possible causes and possible effects, is otherwise also referred to as *correlational research*.

The essential difference between descriptive and experimental research lies in the question as to cause and effect. Based on descriptive research, a causal relationship between cause and effect *cannot* be properly established. Descriptive research might show that there is a correlation between a particular type of nutrition and a longer lifespan. Does this mean that this type of nutrition is the cause of a longer lifespan? This is definitely not necessarily the case: it is also possible that this type of food is mainly consumed by people who are relatively highly educated and wealthy, and who live longer because of these other factors[1]. In order to determine whether there is a causal relationship, we must set up and conduct experimental research.

1.5 Experimental research

Experimental research is characterized by the researcher's systematically manipulating a particular aspect of the circumstances under which a study is conducted (Shadish et al., 2002). The effect arising from this manipulation now becomes central in the research study. For instance, a researcher suspects that a particular new method of teaching will result in better student performance compared to the current teaching method. The researcher wants to test this hypothesis using experimental research. She or he manipulates the type of teaching: some groups of students are taught according to the novel, experimental teaching method, and other groups of students are taught according to

[1] It is even possible that the nutrition habits under study cause people to live *shorter*, but that this negative effect is masked by the stronger positive effects of education and wealth.

the traditional method. The novel teaching method's effect is evaluated by comparing both types of student groups' performance after they have been 'treated' with the old vs. new teaching method.

The advantage of experimental research is that we may usually interpret the research results as the consequence or effect of the experimental manipulation. Because the research systematically controls the study and varies just one aspect of it (in this case, the method of teaching), possible differences between the performance observed in the two categories can only be ascribed to the aspect that has been varied (the method of teaching). Logically speaking, this aspect that was varied is the only thing that could have cause the observed differences. Thus, experimental research is oriented towards evaluating causal relationships.

This reasoning does require that test subjects (or groups of students, as in the example above) are assigned to experimental conditions (in our example, the old or the new method of teaching) at random. This random assignment is the best method to exclude any non-relevant differences between the conditions of treatment. Such an experiment with random assignment of test subjects to conditions is called a *randomized experiment* or *true experiment* (Shadish et al., 2002). To remain with our example: if the researcher had used the old research method only with boys, and the new research method only with girls, then any difference in performance can no longer just be attributed to the manipulated factor (teaching method), but also to a non-manipulated but definitely relevant factor, in this case, the students' sex. Such a possible disruptive factor is called a confound. In Chapter **??**, we will discuss how we can neutralize such confounds by random assignment of test subjects (or groups of students) to experimental conditions, combined with other measures.

There also exists experimental research in which a particular aspect (such as teaching method) is indeed systematically varied, but in which test subjects or groups of students are not randomly assigned to the experimental conditions; this is called *quasi-experimental research* (Shadish et al., 2002). In the example above, this term would be applicable if teaching method were investigated using data from groups of students for which it was not the researcher, but their teacher who determined whether the old or new teaching method would be used. In addition, the teacher's enthusiasm or teaching style might be a confound in this quasi-experiment. We will encounter various examples of quasi-experimental research in the remainder of this textbook.

Within the type of experimental research, we can also make a further division: that between laboratory research and field research. In both types of experimental research, some aspect of reality is manipulated. The difference between both types of research lies in the degree to which the researcher is able to keep under control the various confounds present in reality. In laboratory research, the researcher can very precisely determine under which environmental conditions observations are made, which means that the researcher can keep many possible confounds (such as lighting, temperature, ambient noise, etc.) under control. In field research, this is not the case. When 'out in the field', the re-

searcher is not able to keep all (possibly relevant) aspects of reality fully under control.

Example 1.2: Margot van den Berg and colleagues from the Universities of Utrecht, Ghana and Lomé investigated how multilingual speakers use their languages when they have to name attributes like colour, size, and value in a so-called Director-Matcher task (Van den Berg et al., 2017). In this task, one research participant (the 'director') gave clues to another participant (the 'matcher') to arrange a set of objects in a particular order. This allowed the researchers to collect many instances of attribute words in a short period of time ("Put the yellow car next to the red car, but above the small sandal"). The interactions were recorded, transcribed, en subsequently investigated for language choice, moment of language switch, and type of grammatical construction. In this type of fieldwork, however, various kinds of non-controlled aspects in the environment may influence the sound recordings and, thus, the data, including "clucking chickens, a neighbour who was repairing his motorbike and had to start it every other second while we were trying to record a conversation, pouring rain on top of the aluminium roof of the building where the interviews took place." (Margot van den Berg, personal communication)

Example 1.3: When listening to spoken sentences, we can infer from a test subject's eye movements how these spoken sentences are processed. In a so-called 'visual world' task, listeners are presented with a spoken sentence (e.g., "Bert says that the rabbit has grown"), while they are looking at multiple images on the screen (usually 4 of them, e.g., a sea shell, a peacock, a saw, and a carrot). It turns out that listeners will predominantly be looking at the image associated with the word they are currently mentally processing: when they are processing *rabbit*, they will look at the carrot. A so-called 'eye tracker' device allows researchers to determine the position on the screen that a test subject is looking at (through observation of their pupils). In this way, the researcher can therefore observe which word is mentally processed at which time (Koring et al., 2012). Research of this kind is best conducted in a laboratory, where one can control background noise, lighting, and the position of test subjects' eyes relative to the computer screen.

Both laboratory research and field research have advantages and disadvantages. The great advantage of laboratory research is, of course, the degree to which the researcher can keep all kinds of external matters under control. In a laboratory, the experiment is not likely to be disturbed by a starting engine or a downpour. However, this advantage of laboratory research also forms an important disadvantage, namely: the research takes place in a more or less artificial environment. It is not at all clear to what extent results obtained under artificial circumstances will also be true of everyday life outside the laboratory. Because of this, the latter forms a point to the advantage of field researcher: the research is conducted under circumstances that are natural. However, the disadvantage of field research is that many things can happen in the field that may influence the research results, but remain outside of the researcher's control (see example 1.2). The choice between both types of experimental research that a researcher has to make is obviously strongly guided by their research question. Some questions are better suited to being investigated in laboratory situations, while others are better suited to being investigated field situations (as is illustrated by the examples above).

1.6 Outline of this textbook

This textbook consists of three parts. Part I (Chapter 1 to 7) covers research methods and explains various terms and concepts that are important in designing and setting up a good scientific research study.

In part II (Chapters 8 to 12), we will cover descriptive statistics, and in part III (Capters 13 to 17), we will cover the basic methods of inferential statistics. These two parts are designed to work towards three goals.

Firstly, we would like for you to be able to critically evaluate articles and other reports in which statistical methods of processing and testing hypotheses on data have been used. Secondly, we would like for you to have the knowledge and insight necessary for the most important statistical procedures. Thirdly, these parts on statistics are meant to enable you to perform statistical analysis on your own for your own research, for instance, for your internship or final thesis.

These three goals are ordered by importance. We believe that an adequate and critical interpretation of statistical results and the conclusions that may be connected to these is of great importance to all students. For this reason, part I of this textbook devotes considerable attention to the 'philosophy' or methodology behind the statistical techniques and analyses we will discuss later.

We will also give you instructions on how you can perform these statistical analyses yourself in SPSS (a popular software package for statistical analysis) and in R (a slightly more challenging, but also much more powerful and versatile

software package that has been gaining popularity). For students and employees at Utrecht University, both packages are pre-installed in `MyWorkSpace`. SPSS is available at https://SurfSpot.nl for a small fee. R is freely available at https://www.R-project.org. A brief introduction to R can be found at https://hugoquene.github.io/emlar2020/; Dalgaard (2002) offers a longer introduction.

Chapter 2

Hypothesis testing research

2.1 Introduction

Many empirical studies pursue the goal of establishing connections between (supposed) causes and their (supposed) effects or consequences. The researcher would like to know whether one variable has an influence on another. Their research tests the hypothesis that there is a connection between the supposed cause and the supposed effect (see Table 2.1). The best way to establish such a connection, and, thus, to test this hypothesis, is an experiment. An experiment that has been set up properly and is well executed is the 'gold standard' in many academic disciplines, because it offers significant guarantees concerning the validity of the conclusions drawn from it (see Chapter **??**). Put differently: the outcome of a good experiment forms the strongest possible evidence for a connection between the variables investigated. As we discussed in Chapter 1, there are also many other forms of research, and hypotheses can also be investigated in other ways and according to other paradigms, but we will limit ourselves here to experimental research.

Table 2.1: Possible causes and possible effects.

Domain	Supposed cause	Supposed effect
trade	outside temperature	units of ice cream sold
healthcare	type of treatment	degree of recovery
eduction	method of instruction	performance on test
language	age at which L2 learning satrts	degree of proficiency
education	class size	general performance in school
healthcare	altitude	rate of malaria infection
language	age	speaking rate (speech tempo)

In experimental research, the effect of a variable manipulated by the researcher on some other variable is investigated. The introduction already provided an example of an experimental study. A novel teaching method was tested by dividing students between two groups. One group was taught according to the novel method, while the other group was taught as usual. The researcher hoped and expected that her or his novel teaching method would have a beneficial effect, meaning that it would lead to better student performance.

In hypothesis testing research, it is examined whether the variables investigated are indeed connected to one another in the way expected by the researcher. Two terms play a central role in this definition: 'variables' and 'in the way expected'. Before we consider experimental research in more detail, we will first take a closer look at these terms.

2.2 Variables

What is a variable? Roughly speaking, a variable is a particular kind of property of objects or people: a property that may vary, i.e., take different values. Let us look at two properties of people: how many siblings they have, and whether their mother is a woman or a man. The first property may vary between individuals, and is thus a (between-subject) variable. The second property may not vary: if there is a mother, she will always be a woman by definition [at least, traditionally]. Thus, the second property is not a variable, but a constant.

In our world, almost everything exists in varying quantities, in varying manners, or to various extents. Even a difficult to define property, like a person's popularity within a certain group, may form a variable. This is because we can rank people in a group from most to least popular. There are ample examples of variables:

- regarding *individuals*: their length, their weight, shoe size, speaking rate, number of siblings, number of children, political preference, income, sex, popularity within a group, etc.
- regarding *texts*: the total number of words ('tokens'), the number of unique words ('types'), number of typos, number of sentences, number of signs of interpunction, etc.
- regarding *words*: their frequency of use, number of syllables, number of sounds, grammatical category, etc.
- regarding *objects* such as cars, phones, etc.: their weight, number of components, energy use, price, etc.
- regarding *organizations*: the number of their employees, their postal code, financial turnover, numbers of customers or patients or students, number

of surgeries or transactions performed or number of degrees awarded, type of organization (corporation, non-profit, …), etc.

2.3 Independent and dependent variables

In hypothesis testing research, we distinguish two types of variables: dependent and independent variables. The *independent* variable is whatever is presumed to bring about the supposed effect. The independent variable is the aspect that a research will manipulate in a study. In our example where an experiment is conducted to evaluate the effects of a new teaching method, the teaching method is the independent variable. When we compare performance between the students that were taught using the new method and those whose writing instruction only followed the traditional method, we can see that the independent variable takes on two values. In this case, we can give these two values (also called *levels*) that the independent variable can take the names of "experimental" and "control", or "new" and "old". We might also express the independent variable's values as a number: 1 and 0, respectively. These numbers do not have a numerical interpretation (for instance, we might as well give these values the names 17 and 23, respectively), but are used here solely as arbitrary labels to distinguish between groups. The manipulated variable is called 'independent' because the chosen (manipulated) values of this variable are not dependent on anything else in the study: the researcher is independent in their choice of this variable's values. An independent variable is also called a *factor* or a *predictor*.

The second type of variable is the dependent variable. The *dependent* variable is the variable for which we expect the supposed effect to take place. This means that the independent variable possibly cause an effect on the dependent variable, or: it is presumed that the dependent variable's value depends on the independent variable's value - hence their names. An observed value for the dependent variable is also called a *response* or *score*; oftentimes, the dependent variable itself may also be given these names. In our example where an experiment conducted to evaluate the effect a new teaching method has on students' performance, the student's performance is the dependent variable. Other examples of possible dependent variables include speaking rate, score on a questionnaire, or the rate at which a product is sold (see Table 2.1). In short, any variable could be used as the dependent variable, in principle. It is mainly the research question that determines which dependent variable is chosen, and how it is measured.

This being said, it must be stressed that independent and dependent variables themselves must not be interpreted as 'cause' and 'effect', respectively. This is because the study has as its goal to convincingly demonstrate the existence of a (causal) connection between the independent and the dependent variable. However, Chapter **??** will show us how complex this can be.

The researcher varies the independent variable and observes whether this results

in differences observed in the dependent variable. If the dependent variable's values differ before and after manipulating the independent variable, we may assume that this is an effect that the manipulation has on the independent variable. We may speak of a relationship between both variables. If the dependent variable's value does not differ under the influence of the independent variable's values, then there is no connection between the two variables.

Voorbeeld 2.1: Quené et al. (2012) investigated whether a smile or frown influences how listeners process spoken words. The words were 'pronounce' (synthesized) by a computer in various phonetic variants - specifically, in such a way that these words sounded as if pronounced neutrally, with a smile, or with a frown. Listeners has to classify the words a 'positive' or 'negative' (in meaning) as quickly as possible. In this study, the phonetic variant (neutral, smile, drown) takes the place of the independent variable, and the speed with which listeners give their judgment is the dependent variable.

2.4 Falsification and null hypothesis

The goal of scientific research is to arrive at a coherent collection of "justified true beliefs" (Morton, 2003). This means that a scientific belief must be properly motivated and justified (and must be coherent with other beliefs). How may we arrive at such a proper motivation and justification? For this, we will first refer back to the so-called induction problem discussed by Hume (1739). Hume found that it is logically impossible to generalize a statement from a number of specific cases (the observations in a study) to a general rule (all possible observations in the universe).

We will illustrate the problem inherent in this generalization or induction step with the belief that 'all swans are white'. If I had observed 10 swans that are all white, I might consider this as a motivation for this belief. However, this generalization might be unjustified: perhaps swans also exist in different colours, even if I might not have seen these. The same problem of induction remains even if I had seen 100 or 1000 white swans. However, what if I had seen a single black swan? In that case, I will know immediately and with completely certainty that the belief of all swans' being white is false. This principle is also used in scientific research.

Let us return to our earlier example in which we presumed that a new teaching method will work better than an older teaching method; this belief is called H1.

Let us now set this reasoning on its head, and base ourselves on the complementary belief that the new method is *not* better than the old one[1]; this belief is called the null hypothesis or H0. This belief that 'all methods have an equal effect' is analogous to the belief that 'all swans are white' from the example given in the previous paragraph. How can we then test whether the belief or hypothesis called H0 is true? For this, let us draw a representative sample of students (see Chapter **??**) and randomly assign students to the new or old teaching method (values of the independent variable); we then observe all participating students' performance (dependent variable), following the same protocol in all cases. For the time being, we presume that H0 is true. This means that we expect no difference between the student groups' performance. If, despite this, the students taught by the new method turn out to perform much better than the students taught by the old method, then this observed difference forms the metaphorical black swan: the observed difference (which contradicts H0) makes it unlikely that H0 is true (provided that the study was valid; see Chapter **??** for more on this). Because H0 and H1 exclude each other, this means that it is very likely that H1 is indeed true. And because we based our motivation upon H0 and not H1, sceptics cannot accuse us of being biased: after all, we did try to show that there was indeed no difference between the performance exhibited by the students in each group.

The method just described is called falsification, because we gain knowledge by rejecting (falsifying) hypotheses, and not by accepting (verifying) hypotheses. This method was developed by philosopher of science Karl Popper (Popper, 1935, 1959, 1963). The falsification method has interesting similarities to the theory of evolution. Through variation between individual organisms, some can successfully reproduce, while many others die prematurely and/or do not reproduce. Analogously, some tentative statements cannot be refuted, allowing them to 'survive' and 'reproduce', while many other statements are indeed refuted, through which they 'die'. In the words by Popper (1963) (p.51, italics removed):

> " … to explain (the world) … as far as possible, with the help of laws and explanatory theories …there is no more rational procedure than the method of trial and error — of conjecture and refutation: of boldly proposing theories; of trying our best to show that these are erroneous; and of accepting them tentatively if our critical efforts are unsuccessful."

Thus, a proper scientific statement or theory ought to be falsifiable or refutable or testable (Popper, 1963). In other words, it must be possible to prove this statement or theory wrong. A testable statement's scientific motivation, and, therefore, its plausibility increase with each time this statement proves to be immune to falsification, and with each new set of circumstances under which this

[1] Two beliefs are complementary when they mutually exclude each other, like H1 and H0 in this example.

happens. 'Earth's climate is warming up' is a good example of a statement that is becoming increasingly immune to falsification, and, therefore, is becoming increasingly stronger.

Voorbeeld 2.2: 'All swans are white' and 'Earth's climate is warming up' are falsifiable, and therefore scientifically useful statements. What about the following statements?
a. Gold dissolves in water.
b. Salt dissolves in water.
c. Women talk more than men.
d. Coldplay's music is better than U2's.
e. Coldplay's music sells better than U2's.
f. If a patient rejects a psychoanalyst's reading, then this is a consequence of their resistance to the fact that the psychoanalyst's reading is correct.
g. Global warming is caused by human activity.

2.5 The empirical cycle

So far, we have provided a rather global introduction to experimental research. In this section, we will describe the course of an experimental study in a more systematic way. Throughout the years, various schemata have been devised that describe research in terms of phases. The best known of these schemata is probably the empirical cycle by De Groot (1961).

The empirical cycle distinguishes five phases of research: the observation phase, the induction phase, the deduction phase, the testing phase, and the evaluation phase. In this last phase, any shortcomings and alternative interpretations are formulated, which lead to potential new studies, each of which once again goes through the entire series of phases (hence the name, 'cycle'). We will now look at each of these five phases of research one by one.

2.5.1 observation

In this phase, the researcher constructs a problem. This is to say, the researcher forms an idea of possible relationships between various (theoretical) concepts or constructs. These presumptions will later be worked out into more general hypotheses. Presumptions like these may come about in myriads of different ways – but all require for the researcher to have sufficient curiosity. The researcher

may notice an unusual phenomenon that needs an explanation, e.g., the phenomenon that the ability to hear absolute pitch occurs much often in Chinese musicians than in American ones (Deutsch, 2006). Systematic surveys of scientific publications may also lead to presumptions. Sometimes, it turns out that different studies' results contradict each other, or that there is a clear gap in our knowledge.

Presumptions can also be based on case studies: these are studies in which one or several cases are studied in depth and extensively described. For instance, Piaget developed his theory of children's mental development based on observing his own children during the time he was unemployed. These observations later (when Piaget already had his own laboratory) formed the impetus for many experiments that he used to sharpen and strengthen his theoretical insights.

It is important to realize that purely unbiased and objective observation is not possible. Any observation is influenced by theory or prior knowledge to a greater or smaller extent. If we do not know what to pay attention to, we also cannot observe properly. For instance, those that specialize in the formation of clouds can observe a far greater variety of cloud types than the uninitiated. This means that it is useful to first lay down an explicit theoretical framework, however rudimentary, before making any observations and analysing any facts.

A researcher is prompted by remarkable phenomena, case studies, studying the literature, etc. to arrive at certain presumptions. However, there are no methodological guidelines on how this process should come about: it is a creative process.

2.5.2 induction

During the induction phase, the presumption voiced in the observation phase is generalized. Having started from specific observations, the researcher now formulates a hypothesis that they suspect is valid in general. (**Induction** is the logical step in which a general claim or hypothesis is derived from specific cases: my children (have) learned to talk → all children (can) learn to talk.)

For instance, from the observation made in their own social circle that women speak more than men do (more minutes per day, and more words per day), a researcher may induce a general hypothesis: H1: women talk more than men do (see Example 2.2; this hypothesis may be further restricted as to time and location).

In addition, the hypothesis' empirical content must be clearly described, which is to say: the type or class of observations must be properly described. Are we talking about all women and men? Or just speakers of Dutch (or English)? And what about multilingual speakers? And children that are still acquiring their language? This clearly defined content is needed to test the hypothesis (see the subsection on testing below, and see Chapter **??**).

Finally, a hypothesis also has to be logically coherent: the hypothesis has to be consistent with other theories or hypotheses. If a hypothesis is not logically coherent, it follows by definition that it cannot be unambiguously related to the empirical realm, which means that it is not properly testable. From this, we can conclude that a hypothesis may not have multiple interpretations: within an experiment, a hypothesis, by itself, must predict one single outcome, and no more than one. In general, three types of hypotheses are distinguished (De Groot, 1961):

- Universal-deterministic hypotheses.
  These take the general shape of *all As are B*. For example: all swans are white, all human beings can speak. If a researcher can show for one single A that it is not B, then the hypothesis has, in principle, been falsified. A universal deterministic hypothesis can never be verified: a researcher can only make statements about the cases they have observed or measured. If we are talking about an infinite set, such as: all birds, or all human beings, or all heaters, this may lead to problems. The researcher does not know whether such a set might include a single case for which 'A is not B'; there is one bird that cannot fly, et cetera. Consequently, no statement can be made about these remaining cases, which means that the universal validity of the hypothesis can never be fully 'proven'.

- Deterministic existential hypotheses.
  These take the general shape of *there is some (at least one) A that is B*. For example: there is some swan that is white, there is some human being that can speak, there is some heater that provides warmth. If a researcher can demonstrate that there exists one A that is B, the hypothesis has been verified. However, deterministic existential hypotheses may never be falsified. If we wanted to do that, it would be necessary to investigate all units or individuals in an infinite set for whether they are B, which is exactly what is excluded by the infinite nature of the set. At the same time, this makes it apparent that this type of hypotheses does not lead to generally valid statements, and that their scientific import is not as clear. One could also put it this way: a hypothesis of this type makes no clear predictions for any individual case of A; a given A might be the specific one that is also B, but it might also not be. In this sense, deterministic existential hypotheses do not conform to our criterion of falsifiability.

- Probabilistic hypotheses.
  These take the general shape of *there are relatively more As that are B compared to non-As that are B*. In the behavioural sciences, this is by far the most frequently occurring type of hypothesis.
  For example: there are relatively more women that are talkative compared to men that are talkative. Or: there are relatively more highly performing students for the new teaching method compared to the old teaching method. Or: speech errors occur relatively more often at the beginning

rather than at the end of the word. This does not entail that all women speak more than all men, nor does this entail that all students taught by the new method perform better than all students taught by the old method.

2.5.3 deduction

During this phase, specific predictions are deduced from the generally formulated hypothesis set up in the induction phase. (**Deduction** is the logical step whereby a specific statement or prediction is derived from a more general statement: all children learn to talk → my children (will) learn to talk.)

If we presume (H0) that "women talk more than men", we can make specific predictions for specific samples. For example, if we interviewed 40 female and 40 male school teachers of Dutch, without giving them a time limit, then we predict that the female teachers in this sample will say more than the male teachers in the sample (including the prediction that they will speak a greater number of syllables in the interview).

As explained above (§2.4), most scientific research does not test H1 itself, but its logical counterpart: H0. Therefore, for testing a H1 (in the next phase of the empirical cycle), we use the predictions derived from H0 (!), for instance: "women and men produce equal numbers of syllables in a comparable interview".

In practice, the terms "hypothesis" and "prediction" are often used interchangeably, and we often speak of testing hypotheses. However, according to the above terminology, we do not test the hypotheses, but we test predictions that are derived from those hypotheses.

2.5.4 testing

During this phase, we collect empirical observations and compare these to the worked-out predictions made "under H0", i.e., the predictions made if H0 were to be true. In Chapter **??**, we will talk more about this type of testing. Here, we will merely introduce the general principle. (In addition to the conventional "frequentist" approach described here, we may also test hypotheses and compare models using a newer "Bayesian" approach; however, this latter method of testing is outside the scope of this textbook).

If the observations made are extremely unlikely under H0, there are two possibilities.

- (i) The observations are inadequate, we have observed incorrectly. But if the researcher has carried out rigorous checks on their work, and if they take themselves seriously, this is not likely to be true.

- (ii) The prediction was incorrect, meaning that H0 is possibly incorrect, and should be rejected in favour of H1.

In our example above, we derived from H0 (!) the prediction that, within a sample of 40 male and 40 female teachers, individuals will use the same amount of syllables in a standardized interview. However, we find that men use 4210 syllables on average, while women use 3926 on average (Quené, 2008, p.1112). How likely is this difference if H0 were true, assuming that the observations are correct? This probability is so small, that the researcher rejects H0 (see option (ii) above) and concludes that women and men do *not* speak *equal* amounts of syllables, at least, in this study.

In the example above, the testing phase involves comparing two groups, in this case, men and women. One of these two groups is often a neutral or control group, as we saw in the example given earlier of the new and old teaching methods. Why do researchers often make use of a control group of this kind? Imagine that we had only looked at the group taught by the new method. In the testing phase, we measure students' performance, which is a solid B on average (7 in the Dutch system). Does this mean that the new method is successful? Perhaps it is not: if the students might have gotten an A or A- (8 in the Dutch system) under the old method, the new method would actually be worse, and it would be better not to add this new method to the curriculum. In order to be able to draw a sensible conclusion about this, it is essential to compare the new and old methods between one another. This is the reason why many studies involve components like a neutral condition, null condition, control group, or placebo treatment.

Now that we know this, how can we determine the probability of the observations we made if H0 were to be true? This is often a somewhat complex question, but, for present purposes, we will give a simple example as an illustration: tossing a coin and observing heads or tails. We presume (H0): we are dealing with a fair coin, the probability of heads is 1/2 at each toss. We toss the same coin 10 times, and, miraculously, we observe the outcome of heads all 10 times. The chance of this happening, given that H0 is true, is $P = (1/2)^{10} = 1/1024$. Thus, if H0 were to be true, this outcome would be highly unlikely (even though the outcome is not impossible, since $P > 0$); hence, we reject H0. Therefore, we conclude that the coin most likely is not a fair coin.

This leads us to an important point: when is an outcome unlikely enough for us to reject H0? Which criterion do we use for the probability of the observations made if H0 were to be true? This is the question of the level of significance, i.e., the level of probability at which we decide to reject H0. This level is signified as α. If a study uses a level of significance of $\alpha = 0.05$, then H0 is rejected if the probability of finding these results under H0[2] is smaller than 5%.

[2]More accurately: If the probability to find either these results or other results that would differ even more from those predicted by H0 is smaller than 5%, then H0 is rejected.

In this case, the outcome is so unlikely, that we choose to reject H0 (option (ii) above), i.e., we conclude that H0 is most probably not true.

If we thus reject H0, there is a small chance that we are actually dealing with option (I): H0 is actually true, but the observations happen *by chance* to strongly diverge from the prediction under H0, and H0 is falsely rejected. This is called a Type I error. This type of error can be compared to unjustly sentencing an innocent person, or undeservedly classifying an innocent email message as 'spam'. Most of the time, $\alpha = 0.05$ is used, but other levels of significance are also possible, and sometimes more prudent.

Note that significance is the probability of finding the extreme data that were observed (or data even more extreme than that) given that H0 is true:

$$\text{significance} = P(\text{data}|\text{H0})$$

Most importantly, significance is *not* the probability of H0 being true given these data, $P(\text{H0}|\text{data})$, even though we do encounter this mistake quite often.

Each form of testing also involves the risk of making the opposite mistake, i.e., not rejecting H0 even though it should be rejected. This is called a Type II error: H0 is, in fact, false (meaning that H1 is true), but, nevertheless, H0 is not rejected. This type of mistake can be compared to unjustly acquitting a guilty person, or undeservedly letting through a spam email message (see Table 2.2).

Table 2.2: Possible outcomes of the decision procedure.

Reality	Decision	
	Reject H0	**Maintain H0**
H0 is true (H1 false)	Type I error (α)	correct
H0 is false (H1 true)	correct	Type II error (β)
	Convict defendant	**Acquit defendant**
defendant is innocent (H0)	Type I error	correct
defendant is guilty	correct	Type I error
	Discard message	**Allow message**
message is OK (H0)	Type I error	correct
message is spam	correct	Type II error

If we set the level of significance to a higher value, e.g., $\alpha = .20$, this also means that the chance of rejecting H0 is much higher. In the testing phase, we would reject H0 if the probability of observing these data (or any more extreme data) were smaller than 20%. This would mean that 8 times heads within 10 coin tosses would be enough to reject H0 (i.e., judging the coin as unfair). Thus, more outcomes are possible that lead to rejecting H0. Consequently, this higher level of significance entails a greater risk of a Type 1 error, and, at the same

time, a smaller risk of a Type II error. The balance between the two type of error depends on the exact circumstances under which the study is conducted, and on the consequences that each of the two types of error might have. Which type of error is worse: throwing away an innocent email, or letting a spam message through? The probability of making a Type I error (the level of significance) is controlled by the researcher themselves. The probability of a Type II error depends on three factors and is difficult to gauge. Chapter **??** will discuss this in more detail.

2.5.5 evaluation

At the end of their study, the researcher has to evaluate the results the study yielded: what do they amount to? The question posed here is not merely whether the results favour the theory that was tested. The goal is to provide a critical review of the way in which the data were collected, the steps of reasoning employed, questions of operationalization, any possible alternative explanations, as well as what the results themselves entail. The results must be put in a broader context and discussed. Perhaps the conclusions will also lead to recommendations, for example, recommendations for clinical applications or for educational practice. This is also the appropriate moment to suggest ideas for alternative or follow-up studies.

During this phase, the aim is primarily to interpret the results, a process in which the researcher plays an important and personal role as the one who provides the interpretation. Different researchers may interpret the same results in widely different ways. Finally, in some cases, results will contradict the outcome that was predicted or desired.

2.6 Making choices

Research consists of a sequence of choices: from the inspirational observations during the first phase, to the operational decisions involved in performing the actual study, to interpreting the results during the last stage. Rarely will a researcher be able to make the best decision for every choice point, but they must remain vigilant of the possibility of making a bad decision along the way. The entire study is as strong as the weakest link: the entire study is as good as the worst choice in its sequence of choices. As an illustration, we will provide an overview of the choices a researcher has to make throughout the empirical cycle.

The first choice that has to be made concerns the formulation of the problem. Some relevant questions that the researcher has to answer at that moment include: how do I recognize a certain research question, is research the right choice in this situation, is it possible to research this idea? The best answers to such

questions depend on various factors, such as the researcher's view of humankind and society, any wishes their superiors or sponsors might have, financial and practical (im)possibilities, etc.

The research question does have to be answerable given the methods and means available. However, within this restriction, the research question may relate to any aspect of reality, regardless of whether this aspect is seen as irrelevant or important. There are many examples of research that was initially dismissed as irrelevant, but, nevertheless, did turn out to have scientific value, for instance, a study on the question: "is 'Huh?' a universal word?" (Dingemanse et al., 2013) (Example 1.1). In addition, some ideas that were initially dismissed as false later did turn out to be in accordance with reality. For instance, Galilei's statement that Earth revolved around the Sun once was called unjustified. In short, research questions should not be rejected too soon for being 'useless', 'platitudes', 'irrelevant', or 'trivial'.

If the researcher decides to continue their study, the next step is usually studying the literature. Most research handbooks recommend doing a sizeable amount of reading, but how is an appropriate collection of literature found? Of course, the relevant research literature on the area of knowledge in question must be looked at. Fortunately, these days, there are various resources for finding relevant academic publications. For this, we recommend exploring the pointers and so-called "libguides" offered by the Utrecht University Library (see http://www.uu.nl/library and http://libguides.library.uu.nl/home_en). We would also like to warmly recommend the guide by Sanders (2011), which contains many extremely helpful tips to use when searching for relevant research literature.

During the next phase, the first methodological problems start appearing: the researcher has to formulate the problem more precisely. One important decision that has to be made at that point is whether the problem posed here is actually suited for research (§2.4). For instance, a question like "what is the effect of the age of onset of learning on fluency in a foreign language?" cannot be researched in this form. The question must be specified further. Crucial concepts must be (re)defined: what is the age of onset of learning? What is language fluency? What is an effect? And how do we define a foreign language? How is the population defined? The researcher is confronted with various questions regarding definitions and operationalization: Is the way concepts are defined theoretical, or empirical, or pragmatic in nature? Which instruments are used to measure the various constructs? But also: what degree of complexity should this study have? Practically speaking, would this allow for the entire study be completed? In which way should data be collected? Would it be possible at all to collect the desired data, or might respondents never be able or willing to answer such questions? Is the proposed manipulation ethically sound? How great is the distance between the theoretical construct and the way in which it will be measured? If anything goes wrong during this phase, this will have a direct effect upon the rest of the study.

If a problem has been successfully formulated and operationalized, a further ex-

ploration of the literature follows. This second bout of literature study is much more focussed on the research question that has been worked out by this point, compared to the broad exploration of the literature mentioned earlier. On the grounds of earlier publications, the researcher might reconsider their original formulation of the problem. Not only does one have to look at the literature in terms of theoretical content, but one should also pay attention to examples of how core concepts are operationalized. Have these concepts been properly operationalized, and if there might be different ways of operationalizing them, what is the reason behind these differences? In addition, would it be possible to operationalize the core concepts in such a way that the distance between the concept-as-intended and the concept-as-defined become (even) smaller (§**??**)? The pointers given above with regard to searching for academic literature are useful here, as well. After this, the research is to (once again) reflect upon the purpose of the study. Depending on the problem under consideration, questions such as the following should be asked: does the study contribute to our knowledge within a certain domain, does the study create solutions for known stumbling blocks or problems, or does the study contribute to the potential development of such solutions? Does the research question still cover the original problem (or question) identified by superiors or sponsors? Are the available facilities, funds, and practical circumstances sufficient to conduct the study?

During the next step, the researcher must specify how data will be collected. This is an essential step, which influences the rest of the study; for this reason, we will devote an entire chapter to it (Chapter **??**). What constitutes the population: language users? Students? Bilingual infants? Speech errors involving consonants? Sentences? And what is the best way to draw a representative sample (or samples) from this population (or populations)? What sample size is best? In addition, this phase involves choosing a method of analysis. Moreover, it is advisable to design a plan of analysis at this stage. Which analyses will be performed, what ways of exploring the data are envisioned?

All the choices mentioned so far are not yet sufficient for finishing one's preparations. One must also choose one's instruments: which devices, recording tools, questionnaires, etc., will be used to make observations? Do suitable instruments already exist? If so, are these easily accessible and does the researcher have permission to use them? If not, instruments must be developed first (§**??**). However, in this latter case, the researcher must also take the task upon themselves to first test these instruments: to check whether the data obtained with these instruments conform to the quality standards that are either set by the researcher or that may be generally expected of instruments used in scientific research (in terms of reliability and validity, see Chapters **??** and **??**).

It is only when the instruments, too, have been prepared that the actual empirical study begins: the selected type of data is collected within the selected sample in the selected manner using the selected instruments. During this phase, also, there are various, often practical problems the researcher might encounter. An example from actual practice: three days after a researcher had sent out their

questionnaire by mail, a nationwide mail workers' strike was set in motion and lasted two weeks. Unfortunately, the researcher had also given the respondents two weeks' notice to respond by mail. This means that, once the strike was over, the time frame the subjects were given to respond had already passed. What was the researcher to do? Lacking any alternatives, our protagonist decided to approach each of the 1020 respondents by phone, asking them to fill out the questionnaire regardless and return it at their earliest convenience.

For the researcher who has invested in devising a plan of analysis in advance, now is the time of harvest. Finally, the analyses that were planned can be performed. Unfortunately, reality usually turns out to be much more stubborn than the researcher might have imagined beforehand. Test subjects might give unexpected responses or not follow instructions, presumed correlations turn out to be absent, and unexpected (and undesirable) correlations do turn out to be present to a high degree. Later chapters will be devoted to a deeper exploration of various methods of analysis and problems associated with them.

Finally, the researcher must also report on their study. Without an (adequate) research report, the data are not accessible, and the study might as well *not* have been performed. This is an essential step, which, among other things, involves the question of whether the study may be checked and replicated based on the way it is reported. Usually, research activity is reported in the form of a paper, a research report, or an article in an academic journal. Sometimes, a study is also reported on in a rather more popular journal or magazine, targeted towards an audience broader than just fellow researchers.

This concludes a brief overview of the choices researchers have to make when doing research. Each empirical study consists of a chain of problems, choices, and decisions. The most important choices have been made before the researcher starts collecting data.

Chapter 3

Integrity

3.1 Introduction

Scientific research has brought humanity immeasurably great benefits, such as reliable computing technology, high-quality medical care, and an understanding of languages and cultures that are not our own. All these assets are based on scientifically motivated knowledge. Researchers produce knowledge, whose progress and growth comes about because researchers build upon their predecessors' experience and insights.

Example 3.1*: Sir Isaac Newton wrote of his scientific work: "If I have seen further it is by standing on [the] shoulders of Giants" (in a letter addressed to Robert Hooke, dated 5 Feb 1675)[1]. This metaphor can be traced back to the medieval scholar, Bernard de Chartres: "nos esse quasi nanos gigantum umeris insidentes" [that we are like dwarfs sat on giants' shoulders] compared to scholars from the times of Antiquity. The aforementioned quote by Newton has also become Google Scholar's motto (scholar.google.com).

In this chapter, we will discuss the ethical and moral aspects of scientific research. Science is done by human beings, and requires a well-developed sense

[1] A copy of this letter may be found at https://digitallibrary.hsp.org/index.php/Detail/objects/9792; for some background information, see http://www.bbc.co.uk/worldservice/learningenglish/movingwords/shortlist/newton.shtml.

of judgment on the part of the researchers. The *Netherlands Code of Conduct for Research Integrity* (VSNU, 2018) (https://www.vsnu.nl/en_GB/research-integrity) describes how researchers (and students) are to behave. According to this code of conduct, scientific research and teaching should be based on the following principles:

- honesty,
- diligence,
- transparency,
- independence, and
- responsibility.

The following sections will go over how these principles are to be implemented in our actions during the various phases of scientific research. How are we to set up a study, collect and process data, and report on our study in a way that is honest, diligent, transparent, independent, and responsible? This is something we have to think about even before we start working on our project, which is why these topics are discussed at the beginning of this reader, even though we will also refer to terms and concepts that will be worked out in more detail in subsequent chapters.

3.2 Design

To be sure, scientific research does bring us immeasurably great benefits, but this is balanced by considerable cost. This includes direct expenses, such as setting up and maintaining laboratories, equipment, and technical support, but also researchers' salaries, financial compensation for informants and test subjects, travel expenses for access to libraries, archives, informants, and test subject, etc. These direct expenses are usually subsidized by public funds held by universities and other academic institutions. In addition, there is an indirect cost, which is partially borne by informants and test subjects: time and effort that can no longer be spent on something else, loss of privacy, and possibly other risks we are not yet aware of. One often forgotten type of cost is loss of naïveté: a test subject who has participated in an experiment learns from it, and, because of this, will possibly respond differently in a subsequent experiment (see §**??**, under History). This means that any results obtained in this subsequent experiment will generalize less well to other subjects who have a different history, and have not yet participated in a study.

Given its great cost, research has to be thought through and designed in such a way that its expected benefits are reasonably balanced by its expected cost (Rosenthal and Rosnow, 2008, Ch.3). If the chance that a study will yield valid

conclusions is very low, it is better *not* to go ahead with this study, which will save on both direct expenses and indirect cost.

Example 3.2: Suppose that we would like to examine whether 4-year-old bilingual children might have a cognitive advantage over monolingual children of the same age. Based on earlier research, we expect a different of at least 2 points (on a 10 point scale) between both groups (with a "pooled standard deviation" $s_p = 4$, hence $d = 0.5$, §**??** and §**??**).

We then compare two group of $n = 4$ children each. Even if there were actually a difference of 2 points between both groups (meaning, if the hypothesis were true), this study would still have a mere 51% chance of yielding a significant difference: the power of the experiment is only .51 (Chapter **??**), because the two groups contain so few test subjects. It would be better for the four-year-olds and their parents to do engage in other activities (at school, at home, or at work) instead of participating in this study.

However, if $n = 30$ children would participate in each of the two groups, and if there were indeed a 2 point difference between both groups (meaning, if the research hypothesis were true), then the power of the experiment would be .90. This means that bigger groups lead to a much better chance of confirming our study's hypothesis. This elaborate research design will cost more (for the researchers and the children and their parents), but presumably it will also yield much more: a valid conclusion with great impact on society.

A study's design (see Chapter **??**) has to be as efficient as possible, and the researcher has to start thinking about it at an early stage. First of all, efficiency depends on choices regarding how the independent variables are manipulated. Is there a separate group of test subjects for each condition of the independent variable (meaning we have "between-subjects" conditions, like in example 3.2 above)? In a between-subjects design that involves two groups, we need about $n = (5.6/d)^2$ subjects in each group (for further explanation of this, see Gelman and Hill (2007), and see §**??**). Or are all test subjects involved in all conditions (meaning we have "within-subjects" conditions)? A within-subjects design with two conditions requires only $n = (2.8/d)^2$ subjects in each condition, and the study will therefore also have lower expenses and indirect cost

for a much smaller number of test subjects. In general, this means that, if possible, it is much better to manipulate independent variables within subjects than it is between subjects. However, this is not always possible, firstly because individual characteristics only differ between subjects by definition (for example: female/male sex, multilingual/monolingual youth, aphasia/no aphasia, etc.). Secondly, we must take proper care to recognize effects of so-called transfer between conditions, which threaten our study's validity (for example: experience, learning, fatigue, maturation). We will return to this in §**??**.

Being multilingual or being female are characteristics that may only vary between individuals. But other conditions may also vary within individuals, for instance, the day on which a cognitive measurement is taken. Suppose that we expect a difference of $D = 2$ points between cognitive measurements taken on Monday and on Friday, respectively (with $s = 4$ and $d = 0.5$, see example 3.2). If we manipulate the day of measurement between subjects, meaning we make separate groups for children tested on Monday and those tested on Friday, this entails that we need $n = (5.6/0.5)^2 = 126$ children in each group, yielding a total of $N = 252$. However, if we manipulate the day of measurement within subjects, meaning that we observe each test subject on Monday and also on Friday, this entails that we need a total of just $N = (2.8/05)^2 = 32$ children. The within-subjects design means that far fewer children's routines will need to be disturbed for our cognitive measurements. However, we must be properly aware of learning effects between the first and second measurement, and take appropriate precautionary measures. For instance, we can no longer use the same questionnaires in both conditions.

A study's efficiency also depends on the dependent variable, in particular, on the observations' level of measurement (Chapter **??**), accuracy, and reliability (Chapter **??**). The lower the level of measurement, the lower also the study's efficiency. As accuracy goes down, the study's efficiency also goes down, and more subjects and observations will be needed to be able to draw valid conclusions.

Example 3.3: Suppose that we would like to examine a difference between two within-subjects conditions, and suppose that the actual difference between them is 2 points (which yields $s_D = 4$ and $d = 0.5$, see example 3.2). However, suppose that we decide to look only at the *direction* and not at the size of the difference between the two observations for each subject: does the subject have a positive or negative difference between the first and second condition? This binomial dependent variable contains less information than the original point score (it contains just the direction and not the size of the difference), making the study less efficient. For this specific example, this means we would need 59 instead of 34 test subjects.

Thus, researchers are responsible for diligently and honestly considering and balancing their study's cost and benefits, and they need to have a sufficient methodological background to be able to choose a proper research design, taking in account time constraints, the available test subjects and instruments of measurement, etc.

3.3 Participants and informants

Scientific research is done by human beings: researchers are but human. In the realm of humanities, these researchers themselves study (other) human beings' behaviour and intellectual products. These activities are governed by laws, rules, guidelines, and codes of conduct that researchers (and students!) must follow, stemming from the aforementioned principles of diligence and responsibility. A study and the data collected for it may not lead to any kind of harm or significant loss of privacy for the parties involved.

For research in the humanities in the Netherlands, two laws are relevant:

- The General Data Protection Regulation (GDPR), see https://autoriteitpersoonsgegevens.nl/nl/onderwerpen/avg-europese-privacywetgeving (in Dutch) or https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en,
- Wet Medisch-wetenschappelijk Onderzoek met mensen (WMO; English: Medical Research Involving Human Subjects Act), see https://wetten.overheid.nl/BWBR0009408/2019-04-02 (in Dutch) or https://english.ccmo.nl/investigators/legal-framework-for-medical-scientific-research/laws/medical-research-involving-human-subjects-act-wmo

It is compulsory to ask participants (or their legal guardians) for their explicit informed consent. This means that participants are fairly informed about the study, about its cost and benefits, and about their remuneration, and that, after this, they explicitly consent to participate. For researchers and students at Utrecht University, helpful examples of informed consent (information letters and consent statements) can be found on the website of the Faculty Ethics Review Committee for the Humanities (FETC-GW, discussed in more detail below), via https://fetc-gw.wp.hum.uu.nl.

All data that may be used to identify an individual are considered to be *personal data*, which may only be collected and processed according to the GDPR. It is advisable to separate one's research data from any personal data as early as possible, which means anonymizing the data. Any information that links personal data and research data (e.g., a list with test subjects' names and their corresponding anonymous personal code) is, itself, confidential and must be

saved and stored with care. Do not keep personal data any longer than necessary. Research data may only be used for the (scientific) goal for which they were collected. Make sure that participants and informants are not recognizable in reports and publications on the study (i.e., use anonymous codes).

Photos and recordings of individuals (including audio, video, physiological data, and EEG) are subject to what we call *portrait rights*. This means that photos and other identifying recordings are considered on a par with portraits. When such a photo or recording is published, the person shown or represented may appeal to their portrait rights and claim damages for the harm done to them by this publication. This means that, if you might be interested in publishing a recording from which someone could be recognized, you must ask the individual who was recorded or their legal guardian for explicit consent beforehand (see above for the notion of informed consent). This also applies if you intend to demonstrate or show a fragment of such a recording at a conference presentation or on a website.

The Dutch WMO law (see above) states that any research involving human subjects must first be approved by a special committee; for the Faculty of Humanities at Utrecht University, this is the Medical Ethics Assessment Committee (Medisch-Ethische Toetsingscommissie or METC), which is administered by the University Medical Centre (UMC). This committee assesses whether the possible benefits of a study are reasonably balanced against the costs and possible harm done to test subjects.

Most research in languages and communication at Utrecht University is exempt from review by the METC, which would otherwise be time-consuming, but it must be submitted to the Faculty Ethics Review Committee for the Humanities (Facultaire Ethische Toetscommissie - Geesteswetenschappen or FETC-GW, see https://fetc-gw.wp.hum.uu.nl/en/). However, this does not apply to research done by students, provided that some conditions apply. You can find more information on the FETC-GW website. When in doubt, always consult with your supervisor or teacher. This ethics assessment is also compulsory for students and researchers in other fields (literature, history, media & culture) who plan to do research with human subjects.

3.4 Data

The data collected form the motivation and empirical basis for the conclusion drawn from scientific research. These data therefore have an essential importance: no data means no valid conclusions. Moreover, as we saw above (§**??**), these data are very costly (in terms of time, money, privacy, etc.). This means that we should treat them very diligently. We must be able to convince others of our conclusions' validity based on these data, and we must be able to share the underlying data with other researchers, if asked.

Thus, diligence requires, at the very least, making a sufficient number of backup copies as soon as possible. Think of what might happen if a fire or flood would completely destroy the place where you work or live, or if your laptop would be stolen during your thesis project (this actually happened to one of our students!). If so, would proper and recent copies of the data be stored in other locations? For storing backup copies, a sufficiently secured cloud service[2] is a good option.

Diligence also requires a proper record of what the data stand for, and how they were collected. Data without a matching description are practically useless for scientific research. Charles Darwin carefully noted down which bird found on which of the Galapagos Islands had which beak shape, and these observations later formed (a part of) the motivation for his theory of evolution. In the same way, we strongly encourage you to keep a log (on paper or digitally) of all steps of your research study, including motivations for these steps, if needed. Also note the brand, type, and settings used for any equipment you use, and note the version and settings for any software used. Keep a record of which processing steps were applied to the data, and why, and which file contains which data.

If you are working with digitized data (e.g., in Excel, or SPSS, or R), make sure to carefully keep track of which variable is stored in which column, using which unit of measurement and which coding scheme.

Example 3.4: The file found at <http://tinyurl.com/nj4pjaq. contains data from 80 speakers of Dutch, partially taken from the Corpus of Spoken Dutch (Corpus Gesproken Nederlands or CGN). The first line contains the variable names. Each subsequent line corresponds to one speaker. The pieces of data on each line are separated by spaces. The first column contains the anonymized speaker ID code, as used in the CGN. In the fifth column, the speaker's region of origin is coded with a single letter: `W` for Western region (Randstad), `M` Central (Mid), `N` North, `S` South) (Quené, 2008). Because of the careful annotation, these data may still be used with no problem, even if they were collected over 20 years ago by fellow researchers.

Data remain the intellectual property of those who collected them. Use of other researchers' data with no citation may be seen as theft or plagiarism.

Data fraud (fabricating data, meaning, coming up with data out of thin air, instead of observing them) is obviously at odds with multiple principles in the code of conduct mentioned above (VSNU, 2018). Fraud harms the mutual trust

[2]Students and employees of most Dutch educational institutions can use SurfDrive (https://www.surfdrive.nl) for easy data storage on secured servers.

on which science is based. It misleads other researchers who might be building on the fictional results, and any research funds allotted to a fraudulent line of research are taken away from other, non-fraudulent research – in other words, it is a mortal sin of academia. If you would like to discuss any questions or dilemmas around this topic, please contact prof.dr. Christoph Baumgartner, confidential advisor on academic integrity for the Humanities at Utrecht University (`c.baumgartner@uu.nl`).

3.5 Writing

Scientific research only really reaches its purpose once its results are being divulged. Research that is not reported on could as well *not* have been conducted at all, and the cost associated with this research was, basically, spent in vain. For this reason, reporting research results is an important part of academic work. Publications (as well as patents) form a very important part of the "output" of scientific research. Researchers are measured by the number of their publications and these publications' "impact" (the number of times these publications are cited by others who build upon them). This great importance is one of the reasons we ought to be diligent in treating others' writings, as well as our own.

The researchers involved in a study must discuss amongst each other who will be listed as authors of a report or publication, and in which order. Those listed as co-authors of a research report have to satisfy three conditions (Office of Research Integrity, 2012, Ch.10). Firstly, the must have made a substantial academic contribution to one or more phases of the study: think of the original idea, setting up and designing the study, collecting the data, or analysing and interpreting the data. Secondly, they must have been a part of writing up the report, either by doing part of the writing or by providing comments on it. Thirdly, they must have approved the final version of the report (most often implicitly, sometimes explicitly), and they must also have consented to being a co-author. It is best practice for the researchers to come to a mutual agreement on the order in which their names are listed. Usually, names are ordered by decreasing importance and extent of each author's contribution. If the lead researcher is the main investigator and also a co-author, this person is often listed last.

Example 3.5: A, a student research assistant, helped collect data, but has made no other contributions, and is not entirely sure what the research is about. This means that A need not be listed as a co-author on the report, but the authors do have to describe and acknowledge A's contribution in their report.

> B, another student, conducted one of the parts of the research project supervised by researcher C. Supervisor C thought of the entire project, but B has collected literature, set up and conducted one part of the study, collected, analysed, and interpreted data, and reported on this all in a paper. Because of this, B and C are both co-authors of a publication on B's part of the research project. They come to an agreement on the order in which authors are listed. Because student B was the most prominent person in the work, while C was the main investigator, they agree that B will be first author and C will be second (and last) author.

Researchers build upon their predecessors' work (see example 3.1). This may also involve building upon their arguments and even their writing, but these cases do require that we always correctly refer to the appropriate source, i.e., to these predecessors' work. After all, if we did not do this, we could no longer distinguish who is responsible for which thought or which fragment of writing. Plagiarism is "copying others' documents, thoughts, arguments, and passing them off as one's own work" (Van Dale, 12th edition [our translation]). This form of fraud is also a mortal sin of academia that may lead to substantial sanctions. The Faculty of Humanities at UU has the following to say about it:

> Plagiarism is the appropriation of another author's works, thoughts, or ideas and the representation of such as one's own work. The following are some examples of what may be considered plagiarism:
>
> - Copying and pasting text from digital sources, such as encyclopaedias or digital periodicals, without using quotation marks and referring to the source;
> - Copying and pasting text from the Internet without using quotation marks and referring to the source;
> - Copying information from printed materials, such as books, periodicals or encyclopaedias, without using quotation marks and referring to the source;
> - Using a translation of the texts listed above in one's own work, without using quotation marks and referring to the source;
> - Paraphrasing from the texts listed above without a (clear) reference: paraphrasing must be marked as such (by explicitly linking the text with the original author, either in text or a footnote), ensuring that the impression is not created that the ideas expressed are those of the student;
> - Using another person's imagery, video, audio or test materials without reference and in so doing representing them as one's own work;

- Resubmission of the student's own earlier work without source references, and allowing this to pass for work originally produced for the purpose of the course, unless this is expressly permitted in the course or by the lecturer;
- Using other students' work and representing it as one's own work. If this occurs with the other student's permission, then he or she may be considered an accomplice to the plagiarism;
- When one author of a joint paper commits plagiarism, then all authors involved in that work are accomplices to the plagiarism if they could have known or should have known that the other was committing plagiarism;
- Submitting papers provided by a commercial institution, such as an internet site with summaries or papers, or which have been written by others, regardless of whether the text was provided in exchange for payment.

https://students.uu.nl/en/practical-information/policies-and-procedures/fraud-and-plagiarism

In the case of self-plagiarism, the fragments or writing or thoughts in question are not taken from others, but from one of the authors. There are various schools of thought on self-plagiarism; however, it is advisable to be sure to cite the relevant source if one is to take ideas from one's own work, building on the principles of diligence, reliability, transparency, and responsibility.

A reference or citation is a shortened mention of a source in the body of the text; you might have seen these quite a few times in this syllabus already. At the end of the report or text, a full list of sources follows, which is usually given the heading, "Sources", "Sources consulted", "References", "Literature", or "Bibliography". A mistake in the references may be seen as a form of plagiarism (Universiteitsbibliotheek, Vrije Universiteit Amsterdam, 2015) because the reader is directed towards an incorrect source. For this reason, it is imperative that researchers cite their sources correctly. Various conventions, depending on the area of study, have been developed for this. Usually, instructors will indicate which style or convention is to be used for citing one's sources. In this textbook, we have intended to follow the style described by the American Psychological Association (2010), a style commonly used in the social sciences and some disciplines within the humanities. (For technical reasons, references may deviate slightly from the APA style.)

The rules for citing sources may sometimes be complex. In addition, authors must make sure that the citations in the body of the text correspond to the list of full references at the end. These tasks are best performed by a so-called "reference manager", a program that collects references or citations, and correctly inserts them into the body of the text. An overview of such programs can be found at https://en.wikipedia.org/wiki/Comparison_of_

reference_management_software. In writing this textbook we have used Zotero (https://www.zotero.org), combined with BibTeX (https://www.bibtex.org).

Bibliography

American Psychological Association (2010). *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, D.C., 6th edition.

Dalgaard, P. (2002). *Introductory Statistics with R*. Springer.

De Groot, A. (1961). *Methodologie: Grondslagen van onderzoek en denken in de gedragswetenschappen*. Mouton, 's-Gravenhage.

Deutsch, D. (2006). The enigma of absolute pitch. *Acoustics Today*, 2:11–19.

Dingemanse, M., Torreira, F., and Enfield, N. (2013). Is "huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items. *PLOS One*, 8(11):e78273.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Hume, D. (1739). *A Treatise on Human Nature*.

Kerlinger, F. N. and Lee, H. B. (2000). *Foundations of Behavioral Research*. Harcourt College Publishers, Fort Worth, 4th edition.

Koring, L., Mak, P., and Reuland, E. (2012). The time course of argument reactivation revealed: Using the visual world paradigm. *Cognition*, 123(3):361–379.

MacFarlane, J. (2020). *Pandoc: a universal document converter*.

Morton, A. (2003). *A Guide through the Theory of Knowledge*. Blackwell, Malden, MA, 3e edition.

Office of Research Integrity (2012). Responsible conduct of research training.

Popper, K. (1935). *Logik der Forschung. Zur Erkentnistheorie der modernen Naturwissenschaft*. Julius Springer, Wien.

Popper, K. (1959). *The logic of scientific discovery*. Routledge, London.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge and Kegan Paul, London.

Quené, H., Semin, G. R., and Foroni, F. (2012). Audible smiles and frowns affect speech comprehension. *Speech Communication*, 54(7):917–922.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *Journal of the Acoustical Society of America*, 123(2):1104–1113.

Rosenthal, R. and Rosnow, R. L. (2008). *Essentials of Behavioral Research: Methods and Data Analysis*. McGraw Hill, Boston, 3e edition.

Sanders, E. (2011). *Eerste Hulp bij e-Onderzoek voor studenten in de geesteswetenschappen: Slimmer zoeken, slimmer documenteren*. Early Dutch Books Online.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth, Belmont, CA.

Universiteitsbibliotheek, Vrije Universiteit Amsterdam (2015). Webcursus informatievaardigheden - algemeen - niveau b.

Van den Berg, M., Amuzu, E. K., Essizewa, K., Yevudey, E., and Tagba, K. (2017). Crosslinguistic effects in adjectivization strategies in Suriname, Ghana and Togo. In Cutler, C., Vrzić, Z., and Angermeyer, P., editors, *Language Contact in Africa and the African Diaspora in the Americas: in honor of John V. Singler*, pages 343–362. Benjamins, s.l.

VSNU (2018). Nederlandse gedragscode wetenschappelijke integriteit. Technical report, VSNU.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.18.