THE EVALUATION AND CRITICAL SYNTHESIS OF EMPIRICAL EVIDENCE

Dr. Tony Hak
Associate Professor of Research Methodology
Rotterdam School of Management
The Netherlands

thak@rsm.nl

TABLE OF CONTENTS

A CHECKLIST FOR CRITICAL EVALUATION OF A RESEARCH REPORT

This checklist contains the elements discussed in most research reports, with routine sections such as "Theory" (or "Literature review"), "Methods", "Results", and "Discussion", in that order. Although it is useful to have an evaluation checklist that follows the order of the report that is evaluated, a slightly different order for evaluation is used here, mainly for didactic reasons.

Checklist for evaluation

1. Is this a study of the hypothesis?
    a. Does the unit of analysis in the study correspond with the focal unit of the hypothesis?
    b. Are the independent variable (IV) and dependent variable (DV) formulated in the hypothesis measured in the study?
2. What is the size of the effect that is observed in this study?
    a. Is an effect size reported that reflects the relation between IV and DV?
    b. Is it a non-standardized or standardized effect size?
    c. How precise is the estimation of the effect size or, in other words, what is the confidence interval?
3. How are results presented and interpreted?
    a. Does the report give the relevant information (effect size and confidence interval)?
    b. Does the report give superfluous information (significance test)?
    c. Does the report's interpretation of the study's results depend on significance or confidence interval?
4. What is the study's research strategy?
    a. Does the hypothesis entail a causal claim?
    b. Does the study's research strategy generate evidence that can support a causal claim?
    c. Is the effect size parameter consistent with the research strategy?
5. Which population is studied?
    a. Is the population a part of the theoretical domain?
    b. Is the population exactly defined and are its characteristics specified?
    c. Is the whole population studied ("census") or a sample?
    d. Is the sample a probability sample?
    e. Are there missing cases ("non-response")? How many?
6. How are the IV and DV measured?
    a. Does measurement rely on informants or respondents?
    b. Are informants or respondents trustworthy?
    c. Are measurements valid?
    d. Are measurements reliable?
    e. Are measurements accurate?
7. Are results inferred to the study's population only?
    a. Does the report make claims beyond the studied population (i.e., a larger population; another population; the theoretical domain; the future)?
    b. Is the theoretical domain assumed to be homogeneous or heterogeneous?
    c. Are results compared (or synthesized) with those of other studies?
    d. What is the practical relevance of the observed effect(s)?

Note that this checklist consists of questions to be asked about the study that is reported, and does not yet mention the criteria by which the answers to these questions should be evaluated. These criteria will be discussed in the various chapters of this book.

Chapter 1 discusses (a) the four elements of a hypothesis; (b) the concept of "effect size"; and (c) bad (significance testing) and good (estimation) ways of analysing and presenting observed effects.

Chapter 2 discusses research strategies. If the hypothesis entails a causal claim, then the research strategy should generate results that can be interpreted as evidence for it. (Not all research strategies do this.) Different research strategies generate different types of effect size.

Chapter 3 discusses populations (including sampling and missing cases) and measurement.

Chapter 4 discusses the claims that can be developed based on the results of a study, and how a study's result should be compared and synthesized with those of other studies ("meta-analysis").

CHAPTER 1
HYPOTHESIS AND EFFECT SIZE

This chapter consists of three parts that correspond with questions 1 to 3 of the checklist for the critical evaluation of research reports.

Checklist for evaluation

1. Is this a study of the hypothesis?
    a. Does the unit of analysis in the study correspond with the focal unit of the hypothesis?
    b. Are the independent variable (IV) and dependent variable (DV) formulated in the hypothesis measured in the study?
2. What is the size of the effect that is observed in this study?
    a. Is an effect size reported that reflects the relation between IV and DV?
    b. Is it a non-standardized or standardized effect size?
    c. How precise is the estimation of the effect size or, in other words, what is the confidence interval?
3. How are results presented and interpreted?
    a. Does the report give the relevant information (effect size and confidence interval)?
    b. Does the report give superfluous information (significance test)?
    c. Does the report's interpretation of the study's results depend on significance or confidence interval?

1.1 Theory and hypothesis

Most studies that report empirical evidence regarding a "hypothesis", present themselves as "theory-testing" studies. What is theory-testing, and why would one do it? One form of empirical research is aimed at describing one or more facts regarding some unit of interest, for instance, the proportion of members of a population who adhere to a specific opinion (in an opinion poll); the average productivity in an economic sector (in a report of a national statistical institute); the success rate of a specific type of project (in an annual report); etc. This descriptive aim requires (a) that the unit of interest is specified, a "population" of people in an opinion poll (e.g., the Dutch adult population), or a population of companies in an economic sector (in a report of a national statistical institute), or a population of projects of a specified kind in a company (in an annual report of that company), and (b) that one or more "variables" (a person's opinion; a company's productivity; a project's success or failure) are measured in each member of that population (or in a probability sample from the population). Often, an implicit or explicit aim of such descriptions is to compare, either between populations; between members within a population; or between different points in time. In some populations (or in specific parts of them) the proportion of people that adhere to an opinion is higher than in others. Productivity appears to be higher in some economic type of company than in another type. The success of some types of project is higher than of others. When such a comparative statement is formulated, it is (at least implicitly) assumed that the stated difference (between different populations or between different types of members of a population) is large enough to bother.

If such a comparative statement is empirically found to be true in several descriptive studies, one might begin to believe that the statement represents a "fact", a difference that probably exists in any similar population. Then such comparative descriptions might be presented as "stylized facts" or "empirical generalizations": women tend to have this opinion more than men; companies of this type are more productive than other types; projects of this type tend to be more successful than other types.

"***Theory***" emerges from an empirical generalization as soon as the generalization moves from a summary of what can be seen in empirically observed populations (women tend to have this opinion more than men in each of the observed populations; etc.) to a claim about what will be the case in *not yet observed* populations (so, we expect that women tend to have this opinion more than men in all populations, not only in the ones that have been observed). A generalization of the latter form is a claim, not a summary of empirical observations. A theoretical statement, thus, is a ***claim*** about what might be the case in not yet observed situations (usually populations). The simplest form of theoretical statement is a claim that there is a difference between types of cases (persons, companies, projects, etc.) in all or in a specified set of situations.

The implicit assumption in a theoretical claim about a difference between types of cases is that this difference is large enough to bother. In other words, the ***size*** of the stated difference matters. And if a difference is large enough to be relevant, than usually it is almost always large enough to be subject to attempts to change it. If some types of projects (type A) are more successful than other types (type not-A) to an extent that matters, then this is an invitation to try to change not-A type projects into type A ones.

Hypothesis

A common format of a theoretical statement is "X is associated with Y" indicating that, in general, cases with a higher value of X have a higher (or lower) value of Y. An example is "A larger workforce is more efficient than a smaller one", which normally means "In cases with a larger workforce there will also be a higher efficiency". The fact that the concepts in a theoretical statement, such as X (e.g., size of the workforce) and Y (e.g., efficiency), can have different values (different sizes) means that they are ***variable*** attributes. Formulated in general terms, a theoretical statement predicts a probability of specific values of an attribute (Y) given a specific value of another attribute (X). The fact that the concepts in a theoretical statement are attributes implies that they are attributes of some entity (i.e., a company or business unit in this example). We will call that entity the "**focal unit**" (FU) of the statement.

A theoretical statement formulates relations between the values of attributes of the focal unit. The focal unit should be specified in the theoretical statement, e.g., "*Manufacturing companies* with a larger workforce will also be more efficient". We call such a statement a **hypothesis** if it is a general statement of the association between X and Y in all instances of the specified entity. "Manufacturing companies with a larger workforce will also be more efficient" is a hypothesis if it, in principle, is a claim about all manufacturing companies, at all times, for all types of products, in all economic sectors, all over the world, etc. If this hypothesis is thought to be true for only specific types of manufacturing (e.g., manufacturing of computers) and only for those

units of a company that are directly involved in manufacturing and sales (and not, e.g., the parts that also provide banking services to clients), then this should be specified in the hypothesis, as in: "For business units that manufacture and sell computers: units with a larger workforce will also be more efficient".

A **theory** is a coherent and consistent set of theoretical statements (*hypotheses*) about a *focal unit*, i.e. it is a set of statements about probabilities of the values of attributes of the focal unit given the values of other attributes of that unit.

> *Other examples*
> *1. "A tangible resource-seeking alliance is more likely to deploy high levels of output and process control"*. This hypothesis states that we can better predict the level of output and process control of an alliance if we know its aim than when we do not know it. Or, in the more technical terms of this chapter: "For focal unit "alliance": if the value of attribute X (the aim of the alliance) is "resource-seeking", the value of attribute Y (the extent of its output and process control) will be higher than with other values of X (other aims)".
> *2. "Affective commitment to change is negatively related to turnover intentions"*. This hypothesis states that we can better predict the turnover intentions of employees if we know their affective commitment to change than when we do not know it. Or, in other words: "For focal unit "employee": if the value of attribute X (an employee's level of affective commitment to change) is high, the value of attribute Y (the desire to quit the company) will be lower than with lower values of affective commitment".

It follows that each theory is defined by four aspects: its focal unit, its domain, its concepts (which represent variable attributes of the focal unit), and relations between concepts as specified in hypotheses. Each of these aspects will be discussed here in more detail.

The **focal unit**, i.e. the unit or entity about which the theory formulates statements, can be very different things, such as activities, processes, events, persons, groups, organizations. If, for example, a theory is formulated about "critical success factors of innovation projects", then *innovation project* is the focal unit. Within a theory, the focal unit should not vary. A theory predicts values of attributes of that focal unit, not of other units. A theory about "critical success factors of innovation projects", for instance, is by definition a theory about characteristics of innovation projects, not of other things or entities such as products, companies, teams, etc. A clear specification of the focal unit is very important in the design of a theory-oriented study because this defines the type of entity about which data must be collected. For estimating the size of the effect entailed in the claim "*A tangible resource-seeking alliance is more likely to deploy high levels of output and process control"* data must be collected about alliances, not about other entities. For estimating the size of the effect entailed in the hypothesis "*Affective commitment to change is negatively related to turnover intentions"* data must be collected about employees, not about entrepreneurs, students or companies.

> ***Question 1a*** of the checklist for critical evaluation of a research report refers to this criterion: *Does the unit of analysis in the study correspond with the focal unit of the hypothesis?* If not, then the study has no value for you if you want to find empirical evidence regarding the hypothesis and hence the study can be discarded in your study.

The **domain** of a theory is the universe of the instances of the focal unit ("**cases**") for which the hypotheses of the theory are assumed to be true. The boundaries of this domain should be specified clearly. For instance, if a researcher develops a theory of critical success factors of innovation projects, it must be clearly stated whether it is claimed that this theory applies to all innovation projects, or only to innovation projects of specific types, or only in specific economic sectors, or only in specific regions or countries, or only in specific time periods, etc. Hence the domain might be very generic (e.g. all innovation projects in all economic sectors in the whole world at all times) or quite specific (e.g. limited to innovation projects in a specific economic sector, in a specific geographical area, or of a specific type, in a specific period).

Examples
1. "A tangible resource-seeking alliance is more likely to deploy high levels of output and process control". This hypothesis is a claim about alliances in general. It is not a claim about a specific type of alliance such as alliances in a specific economic sector (e.g., airline alliances) or in specific countries (e.g., US alliances). If this hypothesis is true, then it is true for alliances in all economic sectors and in all countries and at all times. If the claim is formulated from the outset as only applicable to a specific type of alliance, then this should have been specified in the wording of the hypothesis.
2. "Affective commitment to change is negatively related to turnover intentions". This hypothesis is a claim about employees in general. It is not a claim about a specific type of employee (e.g., manual laborers or white collar workers), or about employees in a specific economic sector (e.g., dockworkers or airline pilots) or in specific countries (e.g., the US workforce). If this hypothesis is true, then it is true for employees in all types of jobs, in all economic sectors, in all countries and at all times. If the claim is formulated from the outset as only applicable to a specific type of employee, then this should have been specified in the wording of the hypothesis.

> ***Question 5a*** of the checklist for critical evaluation of a research report (to be discussed in Chapter 3) refers to this criterion: *Is the population a part of the theoretical domain?* If not, then the study has no value for you if you want to find empirical evidence regarding the hypothesis and hence the study can be discarded in your study.

The **concepts** of the theory designate the *variable attributes of the focal unit*. They specify the "topic" of the theory. "This is a theory about (topic) the determinants of success of a project." "This is a theory about (topic) the effects of having a high … in companies." An attribute described by a concept can be absent or present, smaller or larger, etc. For instance, if the research topic is "critical success factors of innovation projects" the *factors* that presumably contribute to success are variable attributes of these projects. In each instance of the focal unit, these factors can be present or absent, or present to a certain extent. Also, *success* is a variable attribute of the focal unit "project", which can be present or absent, or present to a certain extent, in each instance of the focal unit (i.e. in each specific innovation project). The attributes that are designated by the concepts of the theory must be *defined* to allow for the measurement of their value in instances of the focal unit (cases). When the value of a concept is measured in such cases, it is called a **variable**. For instance, in a theory of critical success factors of innovation projects, the concept "project outcome" needs to be defined such that it is clear what counts as a "successful" outcome and what does not. The "factors" must be defined as well, so that we can measure the extent to which each factor is present.

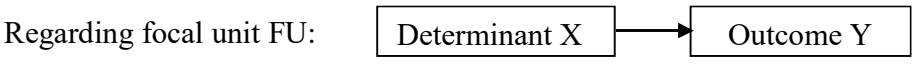
> ***Question 1b*** of the checklist for critical evaluation of a research report refers to this criterion: *Are the independent variable (IV) and dependent variable (DV) formulated in the hypothesis measured in the study?* If not, then the study has no value for you if you want to find empirical evidence regarding the hypothesis and hence the study can be discarded in your study.

The **hypotheses** of a theory formulate relations between the concepts (i.e., between variable attributes) of the focal unit. In the typical case, but not always, this relation is a causal one. A **causal relation** is a relation between two attributes X and Y of a focal unit in which a value of X (or its change) results in a value of Y (or in its change).

> ***Question 4a*** of the checklist for critical evaluation of a research report (to be discussed in Chapter 2) refers to this criterion: *Does the hypothesis entail a causal claim?* If so, then this has implications for what is a good or a less good research strategy.

A hypothesis can be visualized by means of a **conceptual model**. Usually such a conceptual model has inputs (**independent concepts**) on the left hand side and outputs (**dependent concepts**) on the right hand side, linked to each other by arrows that point to the dependent concepts. The arrows are indications of the direction of the causal relation between the concepts. The nature of these arrows needs to be defined more precisely in the wording of the hypotheses of the theory.

The simplest building block of a theory is a single hypothesis that formulates the relation between two concepts. A hypothesis can be visualized as follows:



This simple model visualizes the hypothesis that, in all cases of the focal unit FU, concept X (the determinant or "independent concept") has an effect on concept Y (the outcome or "dependent concept"). The unidirectional arrow represents the assumption that a cause precedes an effect. Because effects are assumed to "depend" on causes, the term "dependent concept" is used for the outcome Y. Causes X are assumed to be "independent" from their effects, hence the term "independent concept".

> Note that this simple model does neither specify the contents of the hypothesis nor the possible values of the concepts. If it is presented in this way, it is normally assumed that X and Y are interval or ratio variables and that the relation between them is causal, probabilistic and positive: "Higher X will on average result in higher Y". Because other types of concepts and other types of relation are possible, it is necessary to add more specifications in the model. "Determinant X (or Y)" should be specified as "Extent of X (or Y)" or "Presence of X (or Y)" or another specification of the (range of) values that are covered by the hypothesis. Furthermore, a sign (+ for positive; – for negative) must be added to the arrow in the model.
>
> Note also that the focal unit (e.g., "innovation project") is *not* depicted in the model itself because the model represents only the variable attributes (concepts) of which the values are linked in the theory, and not the invariable entity about which the theory is formulated. For this reason, the model is prefaced by a statement about the focal unit. The domain is implied.

More complicated conceptual models might depict relations between a larger number of independent concepts X1, X2, X3, etc., and dependent concepts Y1, Y2, Y3, etc. For instance, in a conceptual model of the "critical success factors of innovation projects", the model would depict a number of different factors (X1, X2, X3, etc.) on the left hand side, "outcome" (as defined precisely) on the right hand side, and an arrow originating from each factor pointing to "outcome". Other models might be used to depict more complex relations such as with moderating or mediating concepts.

> Note that the word "theory" is used loosely in the literature and often refers to sets of statements that are not "theory" as defined here. For instance, "theories" such as "the resource based view" or "transaction cost theory" are perspectives, not sets of precise hypotheses with defined concepts.

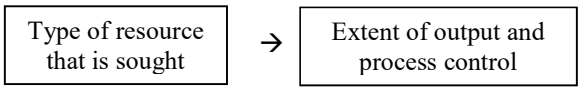
Here follow a number of hypotheses that have been formulated and investigated by Bachelor students in the Research Training & Bachelor Thesis course at the Rotterdam School of Management. For each of these hypotheses, the focal unit, domain, concepts, relations, and conceptual model are specified in the following examples; and conclusions are drawn regarding the unit of analysis and variables in a research report.

Example 1

Hypothesis:	"A tangible resource-seeking alliance is more likely to deploy high levels of output and process control"
Focal unit:	Alliance
Domain:	All alliances in the world, in all economic sectors, in all countries, at all times
Independent concept:	Type of resource that is sought in the alliance (tangible versus intangible)
Dependent concept:	Extent to which output and process control is used
Relation:	Probably causal, probabilistic, positive

Conceptual model:

Regarding focal unit alliance:



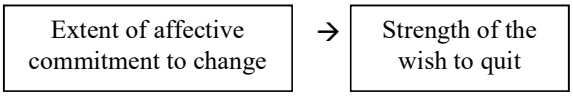
A study of this hypothesis must be a study of a population of *alliances* (or a sample from that population) and must measure for each member of the population (or sample) (a) the type of resource that is sought and (b) the extent of output and process control.

Example 2

Hypothesis:	"Affective commitment to change is negatively related to turnover intentions"
Focal unit:	Employee
Domain:	All employees in the world, in all types of job, in all countries, at all times
Independent concept:	Extent of affective commitment to change (from not at all to very much)
Dependent concept:	Strength of the wish to quit (from not at all to very much)
Relation:	Probably causal, probabilistic, negative

Conceptual model:

Regarding focal unit employee:



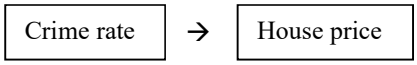
A study of this hypothesis must be a study of a population of *employees* (or a sample from that population) and must measure for each member of the population (or sample) (a) the extent of affective commitment to change and (b) the strength of the wish to quit.

Example 3

Hypothesis:	"Higher crime rates have a negative effect on house prices"
Focal unit:	Neighborhood or city or region
Domain:	All neighborhoods (cities, or regions) in the world, in all countries, at all times, of all types
Independent concept:	Crime rate in the neighbourhood (city, or region)
Dependent concept:	(Average) price of a house in a neighbourhood (city, or region)
Relation:	Causal, probabilistic, negative

Conceptual model:

Regarding focal unit city (etc.):



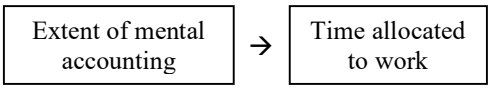
A study of this hypothesis must be a study of a population of *houses* (or a sample from that population) and must measure for each member of the population (or sample) (a) crime rate and (b) the average house price.

Example 4

Hypothesis:	"The way entrepreneurs allocate their time is influenced by their tendency for mental accounting"
Focal unit:	Entrepreneur
Domain:	All entrepreneurs in the world, in all countries, at all times, in all business types
Independent concept:	The extent to which a person evaluates costs and benefits of activities
Dependent concept:	The time allocated to work-related activities (versus other activities such as leisure and family-related) under a time constraint
Relation:	Causal, probabilistic

Conceptual model:

Regarding focal unit entrepreneur:



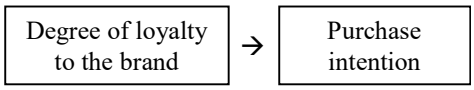
A study of this hypothesis must be a study of a population of *entrepreneurs* (or a sample from that population) and must measure for each member of the population (or sample) (a) the extent of mental accounting and (b) the time allocated to work.

Example 5

Hypothesis:	"Consumers with higher loyalty to a brand respond more favorably to brand extensions in general and to "distant" extensions in particular"
Focal unit:	Consumer
Domain:	All consumers in the world, in all countries, at all times, in all types of consumption
Independent concept:	The degree of loyalty to the brand
Dependent concept:	Level of intention to purchase a (distant) brand extension
Relation:	Causal, probabilistic

Conceptual model:

Regarding focal unit consumer:



A study of this hypothesis must be a study of a population of *consumers* (or a sample from that population) and must measure for each member of the population (or sample) (a) the degree of loyalty to the brand and (b) level of intention to purchase a (distant) brand extension.

You are now able to select research reports that are relevant for your project, in the sense that they potentially contain empirical information about your hypothesis, and distinguish these from reports that do not contain such information because they study another unit of analysis or do not measure both variables in that unit.

In the following it is assumed that you are evaluating a report that has "survived" Question 1 of the checklist. We can now move to question 2.

1.2 Effect size

In the discussion about the four elements of a hypothesis in section 1.1 above, no mention was made of the size of the effect. However, a hypothesis such as "Regarding focal unit FU: X causes Y" is only a shortened version of the theoretical claim that "Regarding focal unit FU: a change in X has an effect on Y ***of a size that matters***". The extent to which an effect *matters* is largely dependent on the managerial or practical implications of the effect.

Managerial relevance

A hypothesis predicts or explains probabilities of values of attributes of a focal unit. Why are such predictions made, and why would you want to know the size of an effect? The obvious answer is that there are many situations in which it matters what the value of an attribute is. For instance, referring to some of the examples above, we would normally prefer lower turnover intentions rather than higher ones; higher house prices rather than lower ones (if we sell or broker); more efficient allocations of time rather than less efficient ones; and higher purchase intentions rather than lower ones. If a researcher develops a theory about a determinant of the extent to which output and process controls are used in an alliance, we expect that this extent matters to people involved in alliances.

If the value of an attribute matters in practice, then it is likely that one would like to be able to manipulate that value. This is the reason that many theories are implicitly, if not explicitly, causal in nature. If the value of one attribute is causally related to the value of another attribute, then it becomes possible (at least in theory) to achieve a higher probability of a desired value of one attribute by manipulating the value of the other attribute. In that sense, causal theories have higher managerial relevance than non-causal theories.

> *Example 2: "Affective commitment to change is negatively related to turnover intentions"*. If this effect is large enough, then it makes managerial sense (to attempt) to increase the average affective commitment to change in a workforce because this might (more likely than not) result in a lower turnover in the company (if a lower turnover is desired).
> *Example 3: "Higher crime rates have a negative effect on house prices"*. If this effect is large enough, then it makes real estate sense (to attempt) to decrease the crime rate in a neighborhood because this might (more likely than not) result in higher house prices in the neighborhood (if a higher house price is desired).

Example 5:"Consumers with higher loyalty to a brand respond more favorably to brand extensions in general and to "distant" extensions in particular". If this effect is large enough, then it makes marketing sense to be careful with brand extensions as long as the average brand loyalty is relatively low (assuming that brands attempt to increase loyalty anyway irrespective of whether they contemplate brand extensions).

Note that the hypothesis in both Examples 1 and 4 in section 1.1 are causal, but that it is not clear how the independent concept could be manipulated in order to achieve an effect in the dependent variable. Managerial relevance is here mainly predictive: "this is what you can expect".

Almost all hypotheses in business research are probabilistic in kind, which means that they do not predict the value of Y in a single case (even if the value of X in that case is known), but only the average value of Y in a set of cases with a specific value of X. This means in practice that the managerial relevance of hypotheses is normally much higher for managers with a portfolio of cases than managers who manage only one or two cases.

Example 2: "Affective commitment to change is negatively related to turnover intentions". This effect, if large enough, is relevant if a manager wants to decrease the average level of turnover in the workforce. It is much less relevant if that manager wants to keep an individual employee with a high value for the company. The latter could be a member of a minority of employees that have a high level of commitment to change and also a high desire to quit.
Example 3: "Higher crime rates have a negative effect on house prices". This effect, if large enough, is relevant if a local government or a real estate developer wants to raise the average house price in a neighborhood. It is much less relevant for individual house owners who want to raise the price of their own homes. Their home could be one of minority of houses that have a low price irrespective of the level of crime in the neighborhood.

Note that the managerial relevance of a hypothesis is dependent on the strength or "effect size" of the relation between the two attributes. If a large increase of affective commitment to change results in only, say, one percentage point decrease in turnover (on average), then the managerial relevance of the hypothesis is doubtful (even if there is some effect). Similarly, if a huge decrease in crime in a neighbourhood results in only, say, one percentage point increase in house prices (on average), then it is doubtful whether the hypothesis is relevant (even if there is such an effect) for a local government or a real estate developer. (Obviously, there might be other good reasons for trying to bring down the crime rate in a neighbourhood.)

Note also that this (more or less subjective) estimation of the practical relevance of a theoretical statement is also dependent on the costs involved in the manipulation of the independent variable. If affective commitment to change could be influenced by a cheap and simple method such as sending an email message to all employees in which they are praised for their efforts, then it is relevant to know that the resulting higher commitment to change has a negative influence (though perhaps small) on the desire to quit. However, if even a small decrease in crime rate requires huge investments in surveillance and other measures, then the effect of such a decrease in the crime rate on house prices should be considerable to make such an investment worthwhile. (Again, obviously, there might be other good reasons for that investment.)

If a hypothesis is considered potentially relevant, and if it is determined how strong the causal effect should be in order to achieve the desired level of relevance, then it becomes useful to conduct studies in which the actual effect size is estimated. For managers it is usually (only) useful to know what the effect size is in the cases which they manage. In contrast, theoreticians and academic researchers are interested in knowing the effect size in the (much larger) theoretical domain.

Effect size measures

Most effect sizes in business research can be grouped into four categories or types:

1. *Measures of association*
   Main examples: regression coefficient; gamma coefficient; and correlation coefficient. These effect size measures can be calculated if both the independent variable and the dependent variable are of a ratio type (or are treated as such, for instance if a nominal variable is treated as a dummy variable in a regression analysis).

2. *Differences between means*
   Main examples: the difference between means of two or more experimental groups; and differences between means of groups defined by the categories of a nominal variable (such as gender, or type of alliance).

3. *Differences between proportions*
   Main examples: abnormal returns in event studies; and differences between proportions of groups defined by the categories of a nominal variable (such as gender, or type of alliance).

4. *Odds ratios and risk ratios*
   Main examples: differences in success or failure between groups.

Note that some (not all) of these effect size measures can be expressed in a non-standardized and in a standardized format. Effect sizes are standardized by expressing them in "scale free" units of the size of one standard deviation as observed in the study. Because standardized measures are "scale free", they are more easily comparable between studies. However, it is more difficult to interpret the practical relevance of a standardized effect size because of the abstract nature of its units. If we want to assess the practical relevance of an observed effect size, it is more intuitive to express the effect in terms of the actual measurement scale.

These measures and their characteristics will be discussed more extensively in Chapter 2. In this phase of this book, we will confine ourselves to identifying the effect size that is reported in the study that you are critically evaluating. This is usually very difficult because most research reports do not clearly report an effect size but something very different: a significance level, usually expressed in terms of *p*-value or t-statistic. Note that *p*-values and t-statistics are *not* effect sizes.

Procedure for finding an effect size in a research report

In most research reports, a number of hypotheses will be formulated in an early part of the report (before the "Methods" section) and a conclusion about these hypotheses will be drawn in a later part of the report (usually the "Results" section). That conclusion is usually based on a number. The problem with routine ways of reporting is that this number is usually not the observed effect size but a significance level. In such cases, our task as a reader is to find the effect size that is observed in the study despite the author's efforts to draw our attention to its "significance".

Procedure for finding the effect size

1. Identify "your" hypothesis in the report. Usually a number of different (though related) hypotheses are "tested" in one study. Usually these hypotheses are numbered (H1, H2, H3a, H3b, etc.). Your hypothesis will be only one of them. Identify the number of your hypothesis.
2. Find in the report the paragraph (or sentence), usually in the "Results" section, in which a conclusion is drawn about the hypothesis. Normally, you can easily identify that paragraph because it mentions the number of your hypothesis (e.g., "H2 is confirmed").
3. Find a reference to a number on which the conclusion regarding the hypothesis is based. If you are lucky, an effect size is mentioned (b=…; r=…; d=…). If you are less lucky, a *p*-value is mentioned or a number of asterisks (such as ***, indicating $p<0.01$). Usually, there is a reference to a table in which relevant numbers are reported, for instance: "Table 4 shows that H2 is confirmed ($p<0.01$)".
4. Find the relevant table; identify the relevant column and row (representing your independent and dependent variables); and find the relevant effect size in the cell defined by your variables.
5. Find out what the relevant effect size measure is. Is the effect size that you have identified in the table a regression coefficient (b=…), a correlation coefficient (r=…), a difference between means (d=…), etc.?
6. Find out whether the effect size is expressed in non-standardized or standardized units.
7. Find (or compute) a confidence interval.

Step 4 in this procedure is crucial and must make sure that you can give a positive answer to Question 2a of the checklist for critical evaluation of a study: *Is an effect size reported that reflects the relation between IV and DV?*

Step 6 in this procedure results in an answer to Question 2b of the checklist: *Is it a non-standardized or standardized effect size?* This is important information for the interpretation of the study's result in terms of practical relevance and for comparison with effect sizes found in other studies.

Step 7 in this procedure results in an answer to Question 2c of the checklist: *How precise is the estimation of the effect size or, in other words, what is the confidence interval?* The width of the confidence interval is an indication of the precision of the point estimate of the effect size.

If you have been successful, you have now answered Question 2 of the checklist for critical evaluation: *What is the size of the effect that is observed in this study?* You must be able to express this effect size in specified units (i.e., it must be clear what the effect size measure is and, if relevant, whether it is non-standardized or standardized) and, if possible, with a confidence interval.

1.3 Interpretation of a study's results

Part 2 of the checklist (What is the size of the effect that is observed in this study?) is not evaluative. After having answered the questions of that part of the checklist, you have an effect size and, in most cases, a confidence interval. You might have had difficulty in finding this information in the research report. In Part 3 of the checklist, you will evaluate the relevance of the information about the effect that is presented and interpreted by the author(s) and the extent to which relevant information is missing. The *general rule* here is that information about the observed effect size and its precision should be foregrounded and that information about "statistical significance", if presented at all, is not used for interpretation. The background of this rule is discussed extensively in Chapters 1 and 2 of Geoff Cumming's book *Understanding the new statistics*.

If the answer to Question 3a of the checklist for critical evaluation of a study (*Does the report give the relevant information: effect size and confidence interval?*) is negative, then it is quite likely that the author will not discuss an effect size and hence will err in drawing conclusions of the results of the study.

If the answer to Question 3b of the checklist (*Does the report give superfluous information: significance test?*) is positive, then it is quite likely that the author will draw conclusions based on "significance" and hence will commit the fallacy of the slippery slope of (non)significance (Cumming, 2012, pp.28-32).

The answer to Question 3c (*Does the report's interpretation of the study's results depend on significance or confidence interval?*) gives usually a pretty good indication of whether the author's conclusions regarding the results of the study can be considered correct and useful. If the report's interpretations depend on the effect size and its confidence interval, then it is not very likely that these interpretations are wrong. However, if these interpretations depend on significance, than it is quite likely that they are (at least partially) wrong. *Note* that we are talking here about the report's *interpretations*, not about the observed *results*. Hence, it is your task as critical reader to "rescue" potentially informative and useful results from an author's misinterpretations due to their application of significance thinking.

CHAPTER 2
RESEARCH STRATEGY

This chapter discusses a study's research strategy, focusing on two main issues:

1. Does the research strategy generate evidence that can support a causal interpretation of the observed effect size? (Question 4b of the checklist for critical evaluation of empirical evidence)
2. Is the effect size parameter consistent with the research strategy (Question 4c of the checklist for critical evaluation of empirical evidence)

A large majority of hypotheses in business research is (implicitly or explicitly) causal in nature. In the literature it is quite common that authors jump from an observed statistically significant result (interpreted as the confirmation of a practically relevant relation between two variables) to an advice to managers to manipulate one variable (the "independent" one) in order to generate a desired effect on another variable (the "dependent" one). Therefore, it is important to make a distinction between types of studies that allow causal inferences and those that do not (or less so). It will be argued that an experimental approach is the golden standard for generating evidence regarding a causal claim.

The term "**research strategy**" will be used here for types of research that differ in (a) the type of claim for which empirical evidence is collected (mainly causal versus non-causal claims) and (b) how such evidence is generated. The main distinction that will be made is between experimental and cross-sectional (non-experimental) research strategies, which results in a typology with three main types: cross-sectional, experimental, and quasi-experimental research strategies. The term "**internal validity**" is introduced to express the extent of fit between the evidence that is generated in a study and the type of claim that the study makes. Internal validity cannot be "measured" but is the outcome of an argumentation. The research strategy that has the lowest level of internal validity for generating evidence regarding a causal effect is the cross-sectional study, which will be discussed first. Then we discuss the experiment, which is the most internally valid research strategy for generating evidence regarding a causal effect. Because, however, there are many causal hypotheses in business research that cannot be experimentally investigated for practical reasons, the question arises whether there are other non-experimental research strategies that have a higher internal validity than cross-sectional research. We discuss some of such non-experimental strategies that are more valid than the cross-sectional study.

In this chapter it is also shown that each type of research strategy can be defined by how the rows ("cases") and columns ("variables") are defined in a corresponding **data matrix**, and that the type ("scale") of the variables determines the type of **effect size parameter** that is most appropriate in the study at hand. This implies that designing an internally valid study consists of the correct specification of the rows and columns of a data matrix. This makes the data matrix the pivot of a research project because the first step of actually conducting a study then consists of filling the cells of the data matrix with scores.

2.1 Associations and the cross-sectional study

A very common research strategy in business research is the cross-sectional study: a "sample" of cases is selected; the independent and dependent variables are measured in the sample; and a multiple regression analysis is applied. This type of study can usually not generate evidence in support of (or against) the causal direction that is implied in the hypothesis (i.e. the claim that one variable has an effect on the other, and not the reverse, or that the effect is caused by another, unobserved or "omitted" variable).

Exercise 2.1

Check whether you fully understand the fact that a cross-sectional study is not an internally valid research strategy for generating evidence regarding a causal claim by answering the following questions.

- Which causal claim can be supported by the evidence that, in the Netherlands, villages with a larger number of cows per inhabitant have also a larger number of children per family?
- Swedish researchers found that marriages in which the husband participates more in household tasks and in the care for the children have a higher divorce rate than marriages with a more conventional division of labour. Critically discuss the conclusion in a Dutch newspaper that husbands' higher rate of participation in household work results in a higher divorce rate.
- Interpret the high correlation between job satisfaction and job performance.

A cross-sectional research strategy is internally valid (only) for a study of the simplest type of hypothesis, namely a study with the aim to assess (only) how much two variables co-vary (i.e., how much change in one variable co-occurs with a change in the value of another variable). We will use the term association for such a co-variation. What type of evidence do we need to support a claim of an association? If both variables in the hypothesis are continuous, we can "observe" such an association in a scatter plot. If the hypothesis expresses a positive association (i.e., it states that, on average, a higher value of one variable is associated with a higher value of the other), then the pattern in the following scatter plot (Figure 2.1) can be seen as supporting evidence. Note that this example is fictitious and is used here only to illustrate in a very introductory way how a relation between two variables can be observed in a scatter plot. (See Cumming, p.382 for much more realistic plots.)

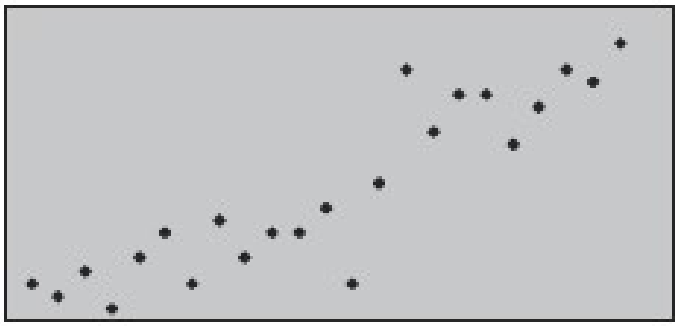


Figure 2.1 Scatter plot showing an association between two variables

Data matrix for a cross-sectional study

Each point in the scatter plot in Figure 2.1 is defined by a value on the horizontal X-axis and a value on the vertical Y-axis. This implies that the plot is derived from a data matrix in which for each case a value for X and a value for Y is specified, as in Figure 2.2.

Cases	Value of X	Value of Y
Case 1	x_1	y_1
Case 2	x_2	y_2
Case 3	x_3	y_3
Case 4	x_4	y_4
Case 5	x_5	y_5
Case 6	x_6	y_6
…	…	…
…	…	…
…	…	…
N =	μ_x	μ_y

Figure 2.2 Data matrix defining a cross-sectional study with continuous variables

There are informal ways by which we can, more or less easily, "see" the association of X and Y in a data matrix. One way of observing an association between X and Y is ranking the cases in the data matrix according to their (increasing or decreasing) value of X and also ranking them according to their (increasing or decreasing) value of Y, and then comparing the two rankings. If the rankings are roughly similar, i.e. if cases are situated high in both rankings (and other cases low in both rankings), this is evidence of an association between X and Y. As shown in Figure 2.1, another way of "seeing" the association between X and Y is plotting the cases in a scatter plot and then observing the empty corners in the plot: top left (low X; high Y) and bottom right (high X; low Y).

The relationship stated in the hypothesis "X is associated with Y" might be notated as a bi-directional connection: X↔Y. In principle, it does not make a difference which variable is called X and which is called Y, and for observing this type of relation (if variables are continuous) it does not matter which variable is on the X-axis and which one is on the Y-axis. A more formal approach than just looking at the scatter plot and seeing a cloud of data points that is stretched from the lower left corner of the plot to its upper right corner would entail plotting a trend line in the scatter plot. However, we can only talk about an "effect size" if we are able to quantify the strength of the association.

Effect size parameters for a cross-sectional study with continuous variables

It is easy to see an association between two variables in a data matrix and even easier to observe it in a scatter plot. However, it is another matter altogether to determine the **strength** of the association (or its "effect size"). One indicator of strength is a value that indicates how much change in the value of Y is associated with how much change in the value of X, on average. The most straightforward indicator of this type of strength is the **regression coefficient**, i.e. the value of *b* in the expression that *describes the trend line*: Y=a+*b*X. A regression coefficient expresses the average change of the value of variable Y (in units of Y) that is associated with a change in the value of variable X (in units of that variable).

An entirely different concept of strength of an association, which requires an entirely different indicator, is the precision with which the value of the one variable in a single case can be predicted when we know the value of the other variable for that case. In practical terms the question is the following. If we know the value of X for a case and compute the value of Y, using the regression formula, how close or far off, on average, would that value be from the observed one? The most straightforward indicator of this type of strength is the **correlation coefficient**. A correlation of 1 corresponds to the situation in which all cases are precisely on a trend line with a positive slope. A correlation of 0 corresponds to the situation in which the scatter plot is a formless cloud through which no trend line can be drawn and hence information about the value of X in a case does not contain any information about its value of Y, i.e. the situation that can be expressed by the expression: Y=a+ε, i.e., a constant with random variation around it.

Note that the non-standardized regression coefficient and the correlation coefficient represent entirely different concepts of strength! A very high correlation can occur together with a very low regression coefficient, e.g. if all data points lie on an almost flat trend line, as in the middle of Figure 2.3.

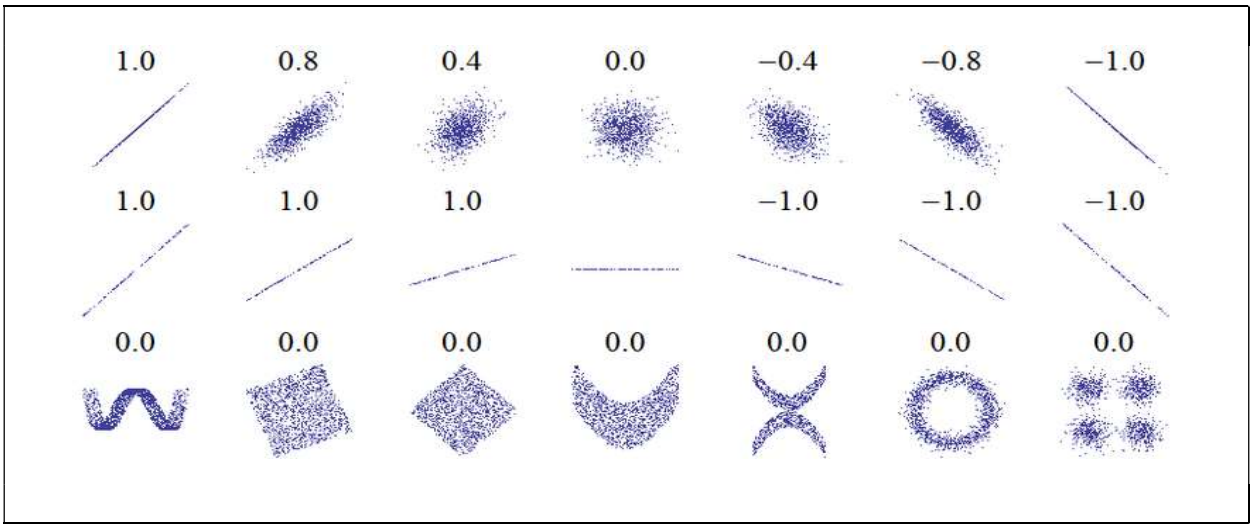


Figure 2.3 Scatter plots that represent different values of correlation.
(Taken from the Wikipedia page on the Pearson product-moment correlation coefficient)

The second row in Figure 2.3 contains a number of trend lines with different slopes (and hence different non-standardized regression coefficients) but all with a perfect correlation of 1. Note that these scatter plots show a distribution of data points in a space defined by X and Y in original scale values and that it is assumed that the X-axis and Y-axis are the same for all plots in this second row of Figure 2.3. Differences in slope, then, indicate equivalent differences in the non-standardized regression coefficients. **Standardized** regression coefficients (for bivariate associations as in this case) are equal to the correlation coefficients, and all are equal to 1 in the second row of Figure 2.3. The explanation for this equivalence is that moving from the left to the middle of this row, the standard deviation of Y is reduced to the same extent as the non-standardized regression coefficient (or slope) is reduced. Hence, the change in Y in *standardized* units stays the same.
In comparison, on the left side of the first row in Figure 2.3 there are three scatter plots with different data clouds (and hence different correlations, as indicated by the number above the plot) but with the same slope. Again, the standardized regression coefficients in this row are equal to the correlations, because the standard deviations of X and Y change accordingly with the increasing scattering of data points (from the left to the middle of the row).
The third row in Figure 2.3 shows relationships in which the linear component is zero and, hence, both coefficients have the value 0.

It is important to note that the correlation coefficient has no causal direction. It is, therefore, not an indicator of the likelihood that the value of one variable will change (in a desired direction) by manipulating the value of the other variable. Correlation has *"predictive"* value: the higher the correlation between X and Y (i.e., irrespective of whether it is a positive or a negative correlation), the more precisely we can predict the value of Y in a case if we know the value of X in that case. That is, if we know the standard deviations of X and Y, because the Pearson product-moment correlation (used here) is a *standardized* measure. The strength of the correlation is the reverse of the average distance from the trend line. However, from the fact that we can "predict" the value of Y if we know the value of X (a prediction which is more or less precise, depending on the strength of the correlation) it does not follow that we can predict the *change* in the value of Y, if any, following a *change* in the value of X. Association or correlation do not entail any information about *causal direction*. In fact, no effect size does contain information about causality. Only research strategies give such information.

A two-fold conclusion follows:

1. An effect size represents the size or strength of a relation, or association.
2. The research strategy in which the effect size was generated allows (or does not allow) to interpret the relation or association as evidence of the strength of a cause-effect relationship.

Exercise 2.2

Take one or more research reports and answer the following questions.

- Is the hypothesis (implicitly or explicitly) intended as a causal one?
- Is the research strategy in the study cross-sectional?
- Develop other interpretations of the study's result than that X causes Y.

We have discussed the logic of cross-sectional research and of correlation and regression with the simple example of a bivariate association (i.e. an association between only two variables). In actual research practice, it is always assumed that a variable of interest Y is associated with a large number of other variables X1, X2, X3, etc., which all should be included in the study. Confusingly, the main argument for the inclusion of multiple variables in a cross-sectional study is usually an argument about causes: Y is supposed to have multiple causes, or a number of "determinants" or "antecedents". This type of causal argument is implicit in the use of multiple regression analysis in the standard research paper. Multiple regression techniques are used routinely in order to "explain" the variance in the value of a "dependent" variable by the variance in the values of multiple "independent" variables.

How does the inclusion of multiple "independent" variables change the study? It hardly does, as long as we analyze each additional variable separately. The data matrix is expanded with a number of columns. These columns (for X1, X2, X3, etc.) are not different in principle from the column for X in the data matrix presented in Figure 2.2. An association between Y and X1, X2, X3, etc., can informally be observed by ranking the cases in the data matrix according to their (increasing or decreasing) value, and then comparing the rankings as in the bivariate case. Also, scatter plots can be made for each association. But things get more complicated if we want to compare the different strengths of these different associations, because the results of a multiple regression analysis cannot be presented in a bivariate scatter plot.

A multiple regression analysis produces a regression coefficient for the regression of Y on each "independent variable" X1, X2, X3, etc., that is "corrected" for the co-variance of the other variables included in the analysis. Because this correction is often formulated as a correction for the "influence" of the other variables, the implicit suggestion of causality is even more easily made than in a bivariate regression analysis. Non-standardized regression coefficients cannot be compared because each "independent" variable has its own scale in which one unit means something radically (and incomparably) different from a unit on the scale of another "independent" variable. For this reason it is quite common to standardize the regression coefficients in a multiple regression analysis. If we compare standardized regression coefficients we are still comparing fundamentally incomparable relations ("apples and pears"; such as the "influence" of the size of a firm on its performance with that of its distance from a port) but we compare them on one common denominator, namely in terms of their observed variance.

However wide the variances of variables might differ, these variances are made equal through standardization. If a standardized regression coefficient for X1 is larger than for X2, this means that a typical variation in the value of X1 explains more of the variation in the value of Y than the typical variation in the value X2 does. Standardized regression coefficients can be transformed into partial correlations in, in principle, the same way as by which a bivariate regression coefficient can be transformed into a correlation coefficient.

Effect size parameters for a cross-sectional study with a nominal variable

If one of the variables in the hypothesis is not continuous, the (informal) ranking of cases and the (more formal) plotting of a trend line are not possible. Take the following hypothesis that was discussed in Chapter 1: "A tangible resource-seeking alliance is more likely to deploy high levels of output and process control". Assume that X and Y have been observed in all members of the population of US airline alliances. Assume also that the population has six members. The data matrix could like look the one in Figure 2.4.

Cases	**Value of X (T = tangible; I = intangible)**	**Value of Y (level of control; scale from 1 to 7)**
Alliance 1	T	6
Alliance 2	T	4
Alliance 3	T	2
Alliance 4	I	5
Alliance 5	I	3
Alliance 6	I	1
N =6	T = 50%; I = 50%	$\mu_y = 3.5$

Figure 2.4 Data matrix defining a cross-sectional study with one discontinuous variable

In this population, an association between the values of X and Y can be observed by comparing the average level of control between the two groups of alliances (tangible resource-seeking and intangible resource-seeking alliances). The observed averages of level of control in this example are 4.0 in the tangible resource-seeking alliances and 3.0 in the intangible resource-seeking alliances. In the analysis, therefore, we would not calculate the trend line but just compute the **difference between the means** of the two groups.

Definition of a cross-sectional study

When we estimate correlation coefficients and regression coefficients, or differences between means, in cross-sectional studies, we assume that cases are "comparable", i.e. that there are no other relevant determinants of X and Y that differ between cases. Because cases in the data matrix must be "comparable" or "similar in relevant respects", all cases in the data matrix should be members of a specified population, i.e. a population in which cases share the characteristics that define it. For instance, in a study of a hypothesis about firms (i.e. the focal unit is a firm), we should generate a data matrix of firms that are members of the same population, e.g. firms in a specific economic sector, in a specific country or region, of a specific size, etc. In other words, the data matrix that we need for a study of an association is the data matrix that is generated in a study of a specific population in which X and Y are observed (or "measured") for each member of that population at one point in time. Hence, we **define** the cross-sectional study as a research strategy in which values of the relevant variables are observed in all members (or in a probability sample of members) of a **population** of instances of a focal unit.

Summarizing this discussion of the analysis of association, the preferred research strategy for studying an association is a cross-sectional study. The values of variables X and Y are observed ("measured") in all members of the population ("cases"), or in a probability sample. The observed values are entered in a data matrix with as many rows as there are cases and with as many columns as there are variables in the hypothesis.
When all variables are continuous, the effect size parameters of interest are the regression coefficient (which, however, easily suggests a causal relation even though such a relation cannot be assessed in this research strategy) and the correlation coefficient. Both coefficients can be generated in a bivariate and multivariate analysis. Only standardized regression coefficients are comparable. Correlation coefficients are always standardized by definition. There is nothing wrong with regression analysis in itself as a tool for describing the mathematical characteristics of a relation between two variables. There is, however, an almost unavoidable – but unjustified – suggestion of causality as soon as the researcher has decided that Y is regressed on X rather than X on Y.
When one of the variables is not continuous, the effect size parameter of interest is usually a difference between means.

2.2 Causal relations and the experiment

An association between two variables might be notated as a bi-directional connection: X↔Y. It does not make a difference which variable is called X and which is called Y, and for observing this type of relation (if variables are continuous) it does not matter which variable is plotted on the X-axis of a scatter plot and which one is on the Y-axis. However, the most common hypothesis is a causal one: "X causes Y" or "Higher values of X cause higher values of Y", X→Y. As suggested by the arrows in the conceptual models of the examples of hypotheses in Chapter 2, each of them is probably meant as describing a causal relation, including those of which the explicit wording only formulates a mere association between X and Y.

A cross-sectional study is not an internally valid research strategy for studying a causal relation. Take, for example, the following hypothesis discussed above: "Consumers with higher loyalty to a brand respond more favorably to brand extensions in general and to "distant" extensions in particular." A non-causal association between these two variables can be observed in a cross-sectional study. Assume that Coca Cola has extended its brand to shoes ("Run to success on Coke!"). A population of Coca Cola consumers could be selected. This could be, for instance, a class of students in business administration. Loyalty to the Coca Cola brand and the intention to buy Coke shoes can be observed (measured) in each of the students in this population and a trend line can be drawn. If this line has an upward slope, this is evidence of an association between brand loyalty and purchase intention. But this association is not necessarily evidence of a *causal* relation. If we want to generate evidence of a causal relation we need a research strategy that allows us to see in the data an *effect* (i.e. a change in Y) that follows a *cause* (i.e. a change in X). This requires an experimental approach. The **experiment** is a research strategy in which the value of an independent variable is manipulated and, next, the value of the dependent variable in each of these instances is observed. The word "next" is crucial because causality implies a temporal order: the effect cannot precede the cause.

Data matrix for an experimental study

An **experiment** is a research strategy in which values of the independent variable are manipulated in instances of the focal unit (recruited from a specified population) and (changes of) values of the dependent variable are observed in these instances. In the most commonly used experimental design, the independent-groups design, the manipulation of the value of the independent variable takes place in two (or more) experimental groups. Cases should be randomly assigned to these groups, then each group gets a "treatment" and, next, it is observed what the value of Y is in these cases. This procedure results in as many *data matrices* as there are groups. Figure 2.5 is an example of two matrices in the simplest form of an independent-groups experiment (with two groups, e.g., one "experimental" and one "control" group). In this example there are two matrices, one for each group that is defined by a value of X (x_1 and x_2) that is assigned experimentally.

Group 1. X= x_1		**Group 2. X=** x_2	
Cases	**Value of Y**	**Cases**	**Value of Y**
Case 1.1	$y_{1.1}$	Case 2.1	$y_{2.1}$
Case 1.2	$y_{1.2}$	Case 2.2	$y_{2.2}$
Case 1.3	$y_{1.3}$	Case 2.3	$y_{2.3}$
Case 1.4	$y_{1.4}$	Case 2.4	$y_{2.4}$
Case 1.5	$y_{1.5}$	Case 2.5	$y_{2.5}$
…	…	…	…
…	…	…	…
…	…	…	…
N =	μ_{x1}	N =	μ_{x2}

Figure 2.5 Data matrices defining a two-independent-groups experimental design

An example

Experimental research is difficult. For instance, it is almost impossible to manipulate someone's loyalty to an existing brand (such as Coca Cola) which might be so much ingrained in a customer that it can even be considered a part of someone's identity. It is also difficult to "control" for other factors than brand loyalty that might influence someone's intention to purchase a brand extension (such as, e.g., negative newspaper reports about the quality of Coke shoes). Students who empirically investigated (a version of) the hypothesis "Consumers with higher loyalty to a brand respond more favorably to brand extensions in general and to "distant" extensions in particular", therefore, invented a fictional brand (Alpha, famous because of its shampoo line) with an equally fictional brand extension (Alpha sporty digital watch). Subjects for the experiment were recruited from the population of students in business administration at the Rotterdam School of Management (N=30). These subjects were randomly assigned to three experimental groups (N=10 each). Each group received a "treatment" which resulted in, respectively, a low, medium and high level of loyalty to the Alpha brand. (Technical details such as how and why this "treatment" works are not discussed here.) Next, each subject was given a description of the Alpha sporty digital watch. Finally subjects were asked how likely it was that they would purchase the watch. This score was entered into the matrix. Assume that the strength of the purchase intention was measured on a scale of 1 to 10. The data matrices could look like the following.

Group 1. Low loyalty		Group 2. Medium loyalty		Group 3. High loyalty	
Cases	**Purchase intention**	**Cases**	**Purchase intention**	**Cases**	**Purchase intention**
Student 1	1	Student 11	3	Student 21	6
Student 2	1	Student 12	3	Student 22	6
Student 3	2	Student 13	4	Student 23	7
Student 4	2	Student 14	4	Student 24	7
Student 5	3	Student 15	5	Student 25	8
Student 6	3	Student 16	5	Student 26	8
Student 7	4	Student 17	6	Student 27	9
Student 8	4	Student 18	6	Student 28	9
Student 9	5	Student 19	7	Student 29	10
Student 10	5	Student 20	7	Student 30	10
N=10	μ_{low}=3	N=10	μ_{medium}=5	N=10	μ_{high}=8

Figure 2.6 Data matrices of a three-independent-groups experimental design

In these three connected matrices, an association between the values of X and Y can be observed: the average level of purchase intention is 3.0 in the low loyalty group, 5.0 in the medium loyalty group and 8.0 in the high loyalty group. This association itself cannot be seen as evidence of a causal relation. However, the fact that none of these subjects could have had a loyalty to the fictional Alpha brand before these subjects had entered the experiment and, hence, no previous intention to purchase an Alpha sporty digital watch could have existed, implies that differences in observed levels of purchase intention can only be "caused" by the experimental treatment, i.e., by the different levels of brand loyalty.

Effect size parameters for an experimental study

If the matrices in Figure 2.6 are filled with scores (or, in other words, if the values of Y have been observed in all cases of both groups), a causal association between X and Y can be observed if the mean value of Y in one group (e.g., the "experimental" group) is different from the mean value of Y in the other group (e.g., the "control" group). Hence, the relevant effect size parameter is the **difference between the means** of the two groups. This is the same effect size parameter as the one discussed above for the cross-sectional study with a discontinuous variable. Here we do not have a discontinuous variable but one that has been manipulated in such a way that two groups of cases have been created such that all members of one group have the same value of the independent variable. The difference between the mean value of Y in the two experimental groups is internally valid evidence of a causal relation because of the preceding manipulation of the value of X, which was absent in the example of two types of alliance.

Note that it is not only assumed that the experimental "treatment" is effective but also that it is equally effective in each member of a group. Without this assumption we could not talk about a difference between groups that differ (only) in the value of the independent variable. It is quite common to test the correctness of this assumption with a "manipulation check". Although such a check yields information about the value of the independent variable in all subjects, this information is normally not used in the analysis.

Figure 2.6 in the example above illustrates that the difference between the mean value of Y (here purchase intention) between groups is the effect size parameter of interest. The observed means of strength of purchase intention is 3.0 in the group with low brand loyalty, 5.0 in the group with medium loyalty and 8.0 in the group with high loyalty. Note that we cannot just calculate the differences between means (2.0 between the low and medium loyalty groups, and 3.0 between the medium and high loyalty groups) but also must construct a confidence interval around these differences. If we would randomly allocate the same 30 students to the three experimental conditions, we would have three groups with another composition than we had in our first experiment and hence we would have slightly different results. In other words, we are confronted with sampling variation and hence we must construct confidence intervals around our results. (See Cumming, Chapter 6, particularly pp.154-164.)

Definition of an experimental study

An **experiment** is a research strategy in which values of the independent variable are manipulated in instances of the focal unit (recruited from a specified population) and (changes of) values of the dependent variable are observed in these cases. An experiment, thus, generates evidence about the size of a causal effect by demonstrating that the effect can be produced (at will) by manipulation of a cause. Experiments have a high level of **internal validity for causal claims** because of this direct link between manipulation and effect, if other potential causes of differences in the values of the dependent variable can be ruled out. In contrast, non-experimental studies have a low internal validity for causal claims because it cannot be known how an observed association came into existence. However, non-experimental studies have generally a higher level of **ecological validity**, because they allow the observation of associations that actually exist (i.e., that are not produced by a researcher's intervention).

Observing a causal relation also requires that other possible influences on the value of the dependent construct (i.e., other causes) are "controlled". This is usually done by designing the experiment in such a way that known influences are the same for each of the cases and that unknown influences are randomly distributed between experimental groups. In order to guarantee the internal validity of the experiment, i.e. to make sure that we really observe an effect of the manipulation of the value of X, cases should be as "comparable" or "similar" as is possible. This is achieved by selecting cases from a specific population (as in the cross-sectional study) and by randomizing these cases over experimental and control conditions.

An additional way of ensuring comparability between the experimental groups is by measuring the value of the dependent variable before the experiment, i.e. by conducting a so-called "pre-test" (which better can be called a "pre-measurement"). If such a pre-test is conducted, the right-hand column in the data matrices does not contain the value of Y ("post-test" or "post-measurement") but the change in the value of Y between pre-test and post-test in each case. The experiment, then, is aimed at generating differences in mean values of the *change in the value of Y* between groups of cases with different values of X.

Note that, in the statistical analysis of experimental results, it is assumed that the cases in the experimental groups are members of a probability sample of a population. This is quite apparent on the **Dance p** page of **ESCI chapters 5-6**, in which the treatment group is considered to be a probability sample of a Treatment population, and the control group is considered to be a probability sample of a Control population. The results of the experiment are interpreted as results for a population. In the large majority of experiments, experimental groups are not probability samples but convenience samples (e.g. students recruited by means of self-selection through an ad in the elevator of a university building). This is a huge problem, because this means that there is no statistical basis for inferring an experimental result to a population, not even to a population of students, and hence that, in a critical synthesis (see Chapter 4 below), we cannot specify a population to which the experimental result can be attributed. This problem is overcome, in principle, in a population-based survey experiment (see Mutz, 2011).

2.3 Quasi-experimental research strategies

An experimental study as outlined above is usually feasible if the focal unit is a person or a situation in which a person can be brought. Studies of psychological theories, for instance, can be conducted by random assignment of people ("subjects") to different experimental conditions. Regrettably, the requirements for conducting a proper experimental study of a hypothesis about *organizations* can rarely be realized in real life, because it is usually difficult to find comparable cases that could be assigned to different experimental conditions. Also the costs of such an experiment (if possible at all) might be prohibitively high. Often it is very costly to intervene in organizations just for the sake of research. Although such obstacles for experiments in real life exist, this does not mean that field experiments, i.e. experiments with companies, business units, projects, and "real" employees, are not possible. This type of experiment is increasingly popular in economic research in general, and in development studies in particular (see Banerjee and Duflo, 2009). Field experiments will not be discussed here because, although getting somewhat more recognition, they are still quite rare in business and management research.

There are, however, a number of alternative research strategies that are not experimental (in the sense that the researcher does not manipulate the value of an independent variable) but have a higher internal validity for causal claims than the cross-sectional study. Here, three of these approaches will be discussed: event study, time series study, and the longitudinal study.

2.3.1 Event study

An *event* is something that is suddenly present and then might cause a change in a dependent variable of interest. An example is the effect of an announcement (e.g. an announcement of a merger) on the stock value of a firm. An **event study** is a research strategy in which changes in the value of a dependent variable are observed in cases in which an event has occurred. Because the value of the dependent variable might have changed anyway even in the absence of the event, the actual change in value of that variable must be benchmarked against what would have happened without the event. In other words, similarly as in an experiment we need a "control" group. But because this is not an experiment in which cases are assigned to the experimental condition (with the event) or the control condition (without the event), the benchmark must be found in some other way. How this is done differs between studies. A common way is to identify a number of similar cases (similar apart from the presence of the event) and to use the average (or weighted) change of value of the dependent variable in those cases as a "control". The most important example of this procedure is the variable "abnormal returns" in event studies regarding stock. An abnormal return is the return (stock value) of a company compared to the average (or weighted) return of similar companies.

An important reason for the similarity between the event study and the experiment is that an event is a discrete phenomenon. The independent variable can have only two values: present or absent. The absence of the event is the normal ("control") condition. The occurrence of the event is the "experimental" condition. Instead of "manipulating" the event (which a researcher cannot do in these cases), the researcher waits for the event to happen. The event study is less internally valid for a causal claim than the experiment, because the possibility cannot be excluded that both the event and the observed change in the value of the dependent variable are caused by an unobserved third variable (another event). The shorter the time span between the event and the change in value of the dependent variable, the more suggestive this is of a causal relation between the two.

Data matrix for an event study

In contrast to the experimental study, in which we have as many data matrices as there are experimental conditions (see Figures 2.5 and 2.6), we have only one data matrix in the event study. This is a data matrix with only the cases with the event. As mentioned above, there are no separate observations of cases without the event because the comparison with those cases is included in the value of Y (stock price, or "return"). For instance, in a study of the effect of an event on stock value, the scores in the data matrix are not the value of Y ("return") for these cases but the difference between this value and the average or weighted value in a benchmark group ("abnormal return").

Figure 2.7 is an example of a data matrix of an event study.

Group with the event. X=1 (present)	
Cases	**Difference of the value of Y with benchmark**
Case 1	y_1
Case 2	y_2
Case 3	y_3
Case 4	y_4
Case 5	y_5
…	…
…	…
…	…
N =	$\mu_{x=1}$

Figure 2.7 Data matrix defining an event study

Effect size parameter for an event study

If the matrix in Figure 2.7 is filled with scores (or, in other words, if the benchmarked values of Y have been observed in all cases of the group with the event), a causal association between X and Y can be observed if the mean value of Y in the cases with the event is different from zero. The value "zero" means that there is no difference between this group and the "control" group. The relevant effect size parameter is the **mean** of Y in the group. If the dependent variable is "abnormal return", then the score is a percentage difference between the case with the event and the comparison group. The effect size in the study is than the average abnormal return, expressed as a percentage.

Definition of an event study

An **event study** is a research strategy in which values of the independent variable (relative to the values in a benchmark group) are observed in instances of the focal unit with the event. An event study suggests causality because the value of the dependent variable is measured immediately after the event. However, the possibility that there is another cause for both the event and the change in the dependent variable cannot be excluded. In that sense, the event study is the poor sister of the experiment.

2.3.2 Time series study

Many relevant independent variables are not discrete events such as announcements, new strategies, disasters, etcetera, but are continuous. We might, for example, be interested in the effect of the size of a specific type of investment (in dollars or euros) on the performance of a firm or a business unit. We cannot conduct an experiment in which we invest more money in one groups of business units than in another group. We do not want to conduct a cross-sectional study either, because that research strategy would not provide us with any evidence about the causal effect. We cannot conduct an event study because we are not interested in the effect of having an investment (however small or large) as compared to not having an investment but, rather, are interested in the effect of the size of the investment. The preferred option, though less attractive than the experiment and the event study (which both are not feasible in this case), is a time series study.

A **time series study** is defined in this book as a research strategy in which it is observed in one instance of the focal unit how the values of the independent and the dependent variables ***change over time***. If changes in the value of the dependent variable are *observably following* the changes in the value of the independent variable, this can be seen as evidence of a causal effect. However, as in the cross-sectional study, there is a possibility that both variables are following changes in the value of a third, unobserved variable. This strategy is much weaker (in terms of internal validity for a causal claim) than the event study because there is no comparison with a "control" group or a benchmark. The internal validity of this research strategy is somewhat higher than the internal validity of the cross-sectional study because of the chronology of (first) the change in the value of the independent variable and (next) the change in the value of the dependent variable. This implies that much care should be given to the choice of the time period (time lag) between the change in the value of the independent variable and the change in the dependent one. Theoretical and practical insights are needed to estimate the time that is needed for an effect to occur. For instance, if we study the effect of the size of an investment on the performance of a business unit, we must know how that investment is supposed to "work", through what mechanisms, etc. Because these mechanisms differ between (types of) business units, field knowledge is essential.

Data matrix for a time series study

In contrast to the event study, in which we have a group of different cases in which the event has occurred, a time series study is a study of only one case. The rows of the data matrix are different points in time or, more precisely, pairs of scores of the independent variable (at one point in time) and of the dependent variable (at a later point in time). For instance, in a study of the effect of the size of investments in new product development and the success in terms of sales of new products, a row in the data matrix will have the investment in certain time period (in the X column) and the sales of new products a number of years later (in the Y column). The number of years is dependent on the average time that it takes for this business (unit) to develop and market a new product. Because we want to stay as close as possible to the experimental design, and hence to the event study, we are interested in how *changes* in the value of X are followed by *changes* in the value of Y. Figure 2.8 is an example of a data matrix of a time series study.

"Cases"=pairs of observations of X and Y separated by time period p	Change in value of X between time 1 and 2, between time 2 and 3, etc.	Change in value of Y between time 1+p and 2+p, between time 2+p and 3+p, etc.
Case 1	x_1	y_1
Case 2	x_2	y_2
Case 3	x_3	y_3
Case 4	x_4	y_4
Case 5	x_5	y_5
Case 6	x_6	y_6
…	…	…
…	…	…
…	…	…
N =	μ_x	μ_y

Figure 2.8 Data matrix defining a time series study

Effect size parameter for a time series study

A causal association between X and Y can be observed if there is a **correlation** between the change in value of X and the later change in the value of Y and if there is a positive **regression** of changes of Y on preceding changes of X. In other words, the effect size parameters for a time series study are the same as for a cross-sectional study. However, a data point in a scatter plot (which corresponds to a row in the data matrix) does not represent the values of X and Y in a case (e.g. a firm) but rather a change in the value of X and a later change in the value of Y in the same instance of the focal unit (e.g. a firm). Different data points in the scatter plot represent different points in time.

Definition of a time series study

A **time series study** is a research strategy in which changes in the value of the independent variable as well as subsequent changes in the dependent variable are observed in one instance of the focal unit. A time series study suggests causality because the change in value of the dependent variable is measured after the change in value of the independent variable. However, the possibility that there is another cause for both changes cannot be excluded. In that sense, the time series study is weaker than the experiment and also weaker than the event study.

The one remaining research strategy that will be discussed in this chapter, the longitudinal study, is even weaker as a strategy for generating evidence about a causal claim.

2.3.3 Longitudinal study

A time series study is only possible if we can collect data for one case over a period of time. That period of time must be much larger than the time needed for an effect to become observable. Let us again take the example of the effect of the size of investment on the success of new product development. If the time needed for the development and marketing of a new product is 5 years and we measure annual changes in X and Y, then we need data for each year over a 16-year period if we want to fill 10 rows in the data matrix. We need data for 11 eleven years (say from year 1 to year 11) in order to observe 10 annual changes in the value of X. Similarly, we need (in this example with an assumed product development period of 5 years) data from year 6 to year 16 in order to observe the corresponding changes in the value of Y. The good news is that we need data for only one instance of the focal unit (e.g. one firm) for one study.

If it is not possible to collect data for many different subsequent points in time in one case (which would mean that a time series study is not feasible), it might be possible to collect data about a limited number of points in time in more cases. Taking again the example of the causal effect of the size of investments on the success of new product development, we might be able to observe a change in investments between two points in time (e.g., between the year 2000 and the year 2003) in the firms of a population and a change in new product success about five years later (e.g., between 2005 and 2008) in the same firms. A **longitudinal study** is a research strategy in which a change in the value of the independent variable and a later change in the value of the dependent variable are observed in a population. If changes in the value of the dependent variable are *associated* with the earlier changes in the value of the independent variable, this can be seen as evidence of a causal effect. However, as in the cross-sectional study, there is a possibility that both variables are following changes in the value of a third, unobserved variable. This strategy is much weaker (in terms of internal validity for a causal claim) than the event study because there is no comparison with a "control" group or a benchmark. It is also weaker than the time series study in this respect because there might be differences in the influence of unobserved variables between the cases of the population, whereas many of such variables do not differ between time points in one firm. In that sense, unobserved variables might be better "controlled" for in a time series study. The internal validity of the longitudinal study (for a causal claim) is still somewhat higher than the internal validity of the cross-sectional study because of the chronology of (first) the change in the value of the independent variable and (next) the change in the value of the dependent variable (as in the time series study).

Data matrix for a longitudinal study

Figure 2.9 is an example of a data matrix of a longitudinal study. This matrix is very similar to the one for a time series study. The columns for the variables are defined in the very same way, namely as a change in the value of X and Y between two time points (of which the change in the value of X precedes the change in the value of Y). However, the cases (rows) are defined very differently. Rows in this data matrix are not pairs of observations at different points in time (in one case). Here rows are separate cases (e.g. firms), all observed at the same points in time ("time 1", "time 2", "time 1+p", and "time 2+p" in Figure 2.9).

Cases	Change in value of X between time 1 and 2	Change in value of Y between time 1+p and 2+p
Case 1	x_1	y_1
Case 2	x_2	y_2
Case 3	x_3	y_3
Case 4	x_4	y_4
Case 5	x_5	y_5
Case 6	x_6	y_6
…	…	…
…	…	…
…	…	…
N =	μ_x	μ_y

Figure 2.9 Data matrix defining a longitudinal study

Effect size parameter for a longitudinal study

As in the time series study, a causal association between X and Y can be observed in a longitudinal study if there is a **correlation** between the change in value of X and the later change in the value of Y and if there is a positive **regression** of changes of Y on preceding changes of X. The effect size parameters for a time series study are, thus, the same as for a cross-sectional study. Data points in a scatter plot represent different members of a population (as in a scatter plot for the cross-sectional study). However, in contrast to the cross-sectional study, these points represent the *change* of the value of X and a *subsequent* change in the value of Y in a case rather than just the values of X and Y at a given point in time.

Definition of a longitudinal study

A **longitudinal study** is a research strategy in which changes in the value of the independent variable as well as subsequent changes in the dependent variable at a later point in time are observed in the members of a population. A longitudinal study suggests causality because the change in value of the dependent variable is measured after the change in value of the independent variable. However, the possibility that there is another cause for both changes cannot be excluded. For this reason, and because there might be unobserved differences between the members of the population, the longitudinal study is weaker in terms of internal validity for a causal claim than all other research strategies except the cross-sectional study.

The most challenging aspect of a longitudinal study is, as with the time series study, to determine how much time should elapse between the change in the value of X and the subsequent change in value of Y.

CHAPTER 3
CASE SELECTION AND MEASUREMENT

This chapter consists of two parts that correspond with questions 5 and 6 of the checklist for the critical evaluation of research reports:

1. Which population is studied? (Question 5 of the checklist for critical evaluation of empirical evidence)
2. How are the IV and DV measured? (Question 6 of the checklist for critical evaluation of empirical evidence)

When a research strategy has been chosen (preferably an experimental study, an event study or a time series study, assuming that the hypothesis is causal), the researcher must construct the corresponding **data matrix** in which each variable has its own column. The research strategy and the data matrix define the effect size parameter that is generated in the study.

> The "Methods" section of a research report will not mention the term "data matrix", but it will mention (a) the research strategy (which, as discussed in Chapter 2 above, determines the general form of the data matrix), (b) the "sample" (which is another word for how the rows of the data matrix are defined) and (c) "measures" is (which is another word for how the cells of the data matrix are filled). Issue (a), research strategy, was discussed in the previous chapter. Issues (b) and (c), case selection and measurement, will be discussed in this chapter.

3.1 Population

We want to estimate an effect that is formulated in a hypothesis or research question. That effect is a relation between characteristics (variables) of a "focal unit", which is preferably explicitly specified in the hypothesis (see Chapter 1). The first step of selection of a "sample" in any study consists of choosing a **population** from the theoretical domain. A population is a set of instances of that focal unit. It is, by definition, a subset of the theoretical domain (which is the total set of all cases to which the hypothesis applies). Figure 3.1 (next page) is a visual representation of the theoretical domain and three of its populations. The domain is populated with all instances of the focal unit ("cases"). In this figure the boundaries of the domain are solid black lines, but in reality the boundaries and the "contents" of the domain will normally be in flux: cases tend to emerge and disappear all the time. The cases in a domain tend to form subsets which can be identified and named, "populations". A population is a set of cases defined by one characteristic or by a set of characteristics. The population of "airline alliances" is an example of a population in the domain of all alliances in the world, in all economic sectors, in all countries, at all times (Example 1 in Chapter 1 above), that is defined by one characteristic (i.e., airline industry) that distinguishes the members of this population from all other alliances in the domain. An example of a population defined by a larger set of characteristics (airline industry; size; region) is the population of "US alliances of airlines with a total turnover of at least [n] dollars a year".

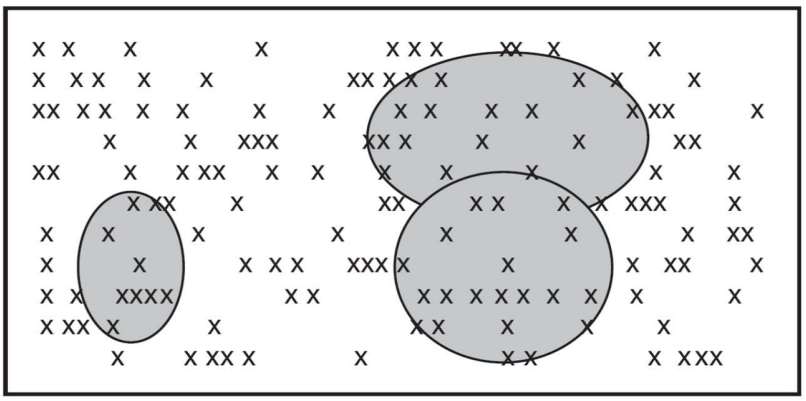


Figure 3.1 Domain, cases and population
(from Dul & Hak, 2008: 46)

The checklist for critical evaluation of empirical evidence contains the following questions regarding "case selection":

5. Which population is studied?
    a. Is the population a part of the theoretical domain?
    b. Is the population exactly defined and are its characteristics specified?
    c. Is the whole population studied ("census") or a sample?
    d. Is the sample a probability sample?
    e. Are there missing cases ("non-response")? How many?

Question 5a (*Is the population a part of the theoretical domain?*) might appear to be superfluous, because it seems so obvious that a population should be selected from the theoretical domain. But this requirement is often violated. For instance, it is quite common that a population of companies or of business units is selected for a study of a hypothesis about projects or teams or, another example, that a population of consumers is selected for a study of a hypothesis about advertisements or brands.

Examples

In a study of a hypothesis about projects (e.g., "More X is associated with more success in projects") it often occurs that a population of companies is selected and that managers of these companies are asked to report whether they think that, in their company, projects with more X are more successful. A proper study of this hypothesis requires that a population of *projects* is selected and that X as well as success is measured for each project.
A common mistake in a study of a hypothesis about brands (such as "Brands with more X have more Y") is to ask consumers whether they think that brands with more X have also more Y. Such a study is an opinion poll, not a proper study of the hypothesis. A proper study requires that a population of *brands* (not consumers) is selected and that X and Y are measured for each brand. It might still be necessary to recruit persons for the experiment. If this occurs, these persons are not cases themselves but only function as raters for the measurement of variables of the brands, etc., such as purchase intention, brand loyalty, etc. The example of such a study discussed in Chapter 2 (see Figure 2.6) demonstrates that usually such a study is a study of only one case (!), Coca Cola in the example, not

of a population of brands. This is usually entirely overlooked by experimenters as well as their readers.

Question 5b (*Is the population exactly defined and are its characteristics specified?*) is important in several respects, both for a critical evaluation of the study itself and for the usefulness of its results for the critical synthesis. Assuming that it is the aim of every single study to generate an effect size for a population, there are only two ways of selecting the cases that are included in the study: either all members of the population are selected ("census") or a sample. The sample must always be a probability sample, for the simple reason that this is an absolute requirement for statistical inference from the sample to the population. In both cases (census and sample), we must have a complete and correct list of members of the population when we conduct a study. We need this list when we conduct a census because all members of the population must be observed (by definition). We also need this list when we draw a probability sample because we can only draw such a sample if we have a complete and correct "sampling frame" (which is the list of members of a population). There are some exceptions to the latter requirement. For instance, in countries that have no population register there is no complete sampling frame for drawing a probability sample. In that case, alternative ways of drawing a sample must be chosen. However, in all cases it is the aim to get a sample that, in statistical respects, can be treated as a probability sample. As discussed above, also cases in experimental groups should be members of a probability sample, in principle. When critically discussing experimental results, it should be taken into account when this requirement is not adhered to (which is often the case).

The problem with results of non-experimental as well as experimental studies in "samples" which are not probability samples is that there is no specific subset of a theoretical domain to which we can infer the study's results. Results pertain only to the (accidental) data set, and cannot be interpreted as "representing" an interpretable category of cases such as a population. This is a problem in a critical synthesis of empirical evidence in a theoretical domain because we cannot fruitfully investigate potential determinants of the variety in effect sizes (for which we need to know the characteristics of the populations in which an effect size is observed). This problem also arises if one wants to draw conclusions about managerial relevance from a study's results. Such conclusions can only be drawn if it is known in which population the effect size is observed, because such conclusions are limited to that population.

Another problem with populations with experiments is that perhaps 90% of experimental studies about hypotheses regarding people or teams (i.e. hypotheses about consumers, employees, managers, etc.) select their cases from only one type of population, namely WEIRD (Western, educated, industrialized, rich, and developed; see Henrich, Heine and Norenzayan (2010).

Question 5c (*Is the whole population studied ("census") or a sample?*) and question 5d (*Is the sample a probability sample?*) are meant to assist in a critical evaluation of sampling procedures, and in drawing a conclusion about the methodological quality of the sample. Convenience sampling (such as "sampling" a list of companies that happen to be known to the researcher, or a university, or a consultancy company, or a training centre, etc.) is very common in business and management research. However, such lists usually form a selection of cases with an unknown relation to the population that they are supposed to "represent". Such lists certainly do not count as "probability" or "random" samples. Hence, a fundamental requirement for the application of inferential statistics is violated in such studies.

Question 5e (*Are there missing cases ("non-response")? How many?*) is meant to assist in a critical evaluation of the data set *after* sampling (or after a population is selected for a census). A requirement for the application of inferential statistics in a data set (i.e., in inferring an effect size in a population from the effect size observed in the data) is that the sample (or the population in a census) is complete. If cases are missing, then we must assume that these cases are not missing at random (MAR) but that some selective (i.e., not random) selection mechanism (NMAR) is playing out. When missing cases are systematically different from observed cases in some way, the observed effect size will be different from the "true" one in the sample or population. The problem with this type of bias ("selection bias" or "non-response bias") is that we do not know how large it is, for the simple reason that the missing cases are missing. The ignorance and negligence about non-response and non-response bias in business and management research is huge. Some repairs have been made after Rogelberg and Stanton published their overview of the problem and of (inevitably deficient) ways by which one might attempt to assess it (Rogelberg & Stanton, 2007).

The general rule here can be simple. If we encounter a study with non-response, the question should be asked how much different could the result have been if those case are substantially different from the ones that have been observed. This is the method of "worst-case resistance", discussed by Rogelberg and Stanton (2007, p.202). The results of this approach, which give some idea of the potential bias resulting from the missingness of cases in the study, are sobering. It will appear that, with the usual response rates in business research, results could have had any other value (i.e., much smaller, much larger, with a reversed sign) than the result obtained in the group of respondents. This is a very good reason for considering the results of any study with more than, say, 20% missing cases as not better than results which would have been obtained by throwing a dice, i.e., useless.

If the simple principles, as outlined in this section 3.1, are applied to mainstream business research, evaluation results are usually very disappointing. Most cross-sectional studies must be considered as *failed studies* because of incorrect sampling procedures and/or non-response. Most experimental studies cannot be considered informative about a population because cases are normally not sampled from a population. On the positive side, missing cases are rare in experiments. It is, however, an open question (because not investigated to a sufficient extent) how experimental results could be biased due to the self-selection of cases (which is the routine selection method).

3.2 Measurement

Although never published in a research report, the core of any study is the data matrix. First, an empty data matrix is constructed that is consistent with the research strategy (see Chapter 2). When a population is selected for the study and when it is decided whether a census or a sample study will be conducted, all cases of the population or sample will be allocated to a row in that data matrix. The number of rows is fixed from then on. There will be as many columns as there will be variables. The design of a study is almost complete at that point, except for one issue, namely a decision about how the cells of the data matrix should be filled. This is a question that must be answered for every column (variable). The technical term for filling a cell in the data matrix is "measurement".

In order to fill the cells of the data matrix we must "measure" the value of each variable in each case. A **variable** is a concept (mentioned in the hypothesis) that is more precisely specified for the cases in the study. This specification must describe in detail the possible values that the variable can have. For instance, if the concept that must be measured is "project success", the variable "success" might be specified in different ways, such as "monetary success" expressed as the amount of dollars that is generated by the project (i.e., a ratio type variable), or as "satisfaction" expressed as a ("subjective") rating by the company on a scale from "none" via "a bit" and "quite a lot" to "very much" (i.e., an ordinal type variable). This specification must be **valid** and the resulting "score" must be **reliable** and **accurate**.

As discussed above, research strategies are defined by their different types of data matrices, not by their methods of measurement. The data that must populate a data matrix can be collected in any (appropriate) way, such as through observations, content analysis, semi-structured interviews and also questionnaires. The latter are only appropriate if data must be collected regarding a person's opinions and beliefs.

Evidence, data and scores

When you fill a cell in the data matrix, usually you cannot just observe some characteristic of the case on that row but must process your observation in some systematic way in order to produce a score that is comparable between cases. It is useful to make a distinction between the following stages of this process: collecting relevant evidence, storing it for further processing, and finally generating a score.
Evidence is what the researcher observes. It can be captured in different ways: in memory, in an audio recording device, as a picture, as a video, as a copy of a document, as a note in a notebook.
Data is evidence that is stored by the researcher in a cupboard or, nowadays more likely, on a hard drive or memory stick. Data are ordered in files in such a way that they can easily be accessed for further processing, but also for audit purposes if someone asks you about the evidence on which your results are based. (Note that, in response to cases of fraud, academic institutions have developed new regulations for the storage of data.) Data are available to the researcher at will (for, e.g., checking or coding). They are not yet a **score**, which is the value of the variable that will be entered into the data matrix. Data must first be categorized or **coded**.

Example

Take the following example of the measurement of a person's opinions or attitudes by means of a semi-structured interview. The interview *evidence* (i.e., what a respondent has said) is recorded in some form (i.e., through a voice recorder). What is recorded might then be transcribed or summarized in a document. This transcript or summary (with the voice recording as a backup and as a source of information about tone of voice, etc.) is the *data*. These data must be interpreted and categorized or *coded* in order to typify the opinions that have been measured in that respondent.

Usually the researcher knows from the outset what kind of score on what kind of "scale" is required or desired as the outcome of the measurement, and this knowledge will steer decisions in the development of the measurement protocol. This can be illustrated with an example of different indicators of a new product development project's success.

Example: three operationalizations of "project success"

Financial success: Usually the score (for the data matrix) should be an amount in some currency (dollar, euro, etc.). Relevant *evidence* for financial success is the financial data about the project. This evidence becomes *data* if it is copied from the company's financial information system (or a report) and is stored in the researcher's data base. If the variable is defined as the *extent* of financial success, a figure can be copied from the data base to the data matrix. If, however, the scale is defined differently, e.g. as the *presence or absence* of financial success, then the dollar or euro amounts must be coded into one of these two possible scores. This means that a coding procedure must be applied by which monetary amounts are evaluated as indicating success (presence/absence), which requires that a cut-off point is specified.
Timely delivery: A criterion must be specified for evaluating the date of delivery as "on time" or "too late" (or any other score deemed relevant for the hypothesis). *Evidence* regarding the date of product delivery as well as the deadline that was stated before must be collected from the company's information system and stored in the researcher's system. Now it is *data*. A *score* for timeliness results from the comparison between these two dates (deadline, and actual delivery).
Satisfaction: Ideally, the company has generated one or more documents in which the project is evaluated. Relevant parts of these documents are the *evidence*. "Text analysis", "document analysis", and "content analysis" are the terms used for generating *scores* from texts. Coding is simple if an evaluation report has a clear conclusion in which the project is unequivocally judged as a success or not. But coding is more complicated if such a judgment must be generated by the researcher from different, ambiguous, and sometimes contradictory, statements in a report. Then the researcher must have a procedure for finding the evaluation result in the text. (Here it is difficult to distinguish a separate phase between evidence and score which could be considered "data".)

These examples illustrate that good measurement requires a lot of decisions which will influence the quality of the evidence first, then the data, and finally the scores that will be analyzed in order to generate the effect size of the study. These decisions must be reported in the research report, so readers can assess measurement quality. When we want to evaluate the quality of a measurement procedure (for every variable separately), we must use three main criteria: measurement validity, measurement reliability, and accuracy.

Measurement validity

A measurement of an attribute (or a variable attribute) is valid if variations in the attribute causally produce variation in the measurement outcomes (Borsboom et al., 2004). It is not possible to "objectively" assess the degree to which measurement validity has been achieved. A level of validity is an outcome of argumentation and discussion. This can be illustrated with the three indicators of success of a project.

Example: three operationalizations of "project success"

Financial success: If financial records must be read in order to retrieve financial evidence relevant for assessing the degree of financial success of a project, be it directly or indirectly (after some computation), the type of financial data that are needed must be precisely specified. It is not possible just to copy any financial number from records but only those numbers whose "meaning" are precisely defined. The "meaning" of a specific number (most often an amount in, say, dollars) is known if it is known how it was produced. If, for instance, the costs involved in a project must be computed from financial records (in order to assess whether a financial gain has occurred), it must be known how the company assigns costs to projects. If relevant costs are not included in the costs documented in the financial records, or when revenues are attributed to the project that actually were generated in ways that are not connected to the project, then the financial success of the project is overestimated. And, reversely, if costs are attributed to the project that actually are not related to the project, or if not all revenue from the project is included in the revenue as documented in the records, then an underestimation of the project's financial success is likely. If necessary, financial data must be recalculated in such a way that they exactly represent *the researcher's definition of the variable*. This must be done for each case in which it is appropriate, for the simple reason that the eventual scores as entered in the data matrix must be comparable. If the records or reports do not contain sufficient information on how the various numbers or amounts have been calculated, it may be necessary to retrieve this information from (financial) staff in order to judge the validity of the evidence. If the evidence is not valid in terms of the researcher's definition, staff could be asked to identify and retrieve other, more valid evidence. In sum, a valid way of extracting evidence of the financial success of a project consists of:

1. Precisely defining what the researcher considers to be the financial success of a project.
2. Translating that definition into precisely described operational procedures.
3. Evaluating the firm's procedures for computing the financial success of a project, if any, against these procedures.
4. If necessary, identifying or computing other, more valid evidence.

The criterion for measurement **validity** of this instrument is whether every detail of its procedures can be justified in terms of the researcher's definition of financial success.

Delivery time: There might be different types of delivery time of project results (the publication of the written report, the oral presentation of the results to management, the final financial record, etc.), of which some might not count as the delivery time as meant in the researcher's definition. Therefore, the researcher must define in a quite detailed way what is considered the delivery time as meant in the hypothesis (and what is not). The researcher's definition must be translated into precise procedures that are applied to candidate pieces of evidence of delivery time which are identified in reading the relevant documents or in the verbal reports from company staff who were involved in the end phase of the project. The criterion for measurement **validity** of these procedures is again whether they are justified in terms of the researcher's definition of delivery time.

Satisfaction: This indicator of success refers to success as defined by the company, not by the researcher. This is an important distinction, which implies that it is not necessary to apply the procedures outlined in the two previous examples. There is no need to evaluate the "correctness" of the company's judgment. The outcome of the company's evaluation can be accepted, irrespective of how it was generated (although the researcher might be interested in the company's procedures and might want to try to collect evidence on these procedures as well). Measurement **validity** here refers to the validity by which the researcher identifies, retrieves, and codes the company's evaluation, irrespective of how the company has generated its evaluation. In case this evaluation has not been recorded in a document by the company, the researcher must (re)construct a company's satisfaction with a project through interviews. There are more and less valid ways of retrieving judgments (such as these evaluations of project success) from respondents in interviews and/or through questionnaires, which will not be discussed here.

Measurement validity, thus, concerns the quality of each element of the measurement protocol or "instrument" in terms of the criteria that follow from the (as precise as possible) definition of the concept that is measured.

Some household examples: A set of scales will be judged as giving a valid measurement of weight because it is clear and transparent, in principle, how weight translates into a specific position of a dial and how different weights give different results on that scale. This applies in the same way to a set of scales for body weight as well as for kitchen scales. It is equally clear how a thermometer translates differences in temperature to different temperature readings. For mercury thermometers this translation follows a different path than for digital thermometers.

It is obvious why we cannot read a temperature on scales and cannot read a weight on a thermometer. However, in business research sometimes a similar mistake is made: a weight is measured as if it is a temperature. This can be an error made by the researcher. It can also be a mistake made by an informant. One common way to achieve efficiency in business research is to use *informants* for measurement. Usually, a questionnaire is sent to such informants. Informants who take the effort to answer the questionnaire must either conduct some sort of measurement (e.g., must weigh something), or must remember it in some way, or must find it in documents or an information system (of weightings performed in the past by the organization), and they must report that information in the questionnaire. In other words, *the informant is asked to conduct the measurement for the researcher without being instructed as a researcher* and, therefore, without knowing the researcher's definition of the variable. The scores obtained in this way usually have questionable validity: the researcher does not know what actually was "measured" by the informant.

Reliability

If a measurement is valid, a next quality criterion is that it is also reliable. **Reliability** is the precision of the scores obtained by a measurement. Reliability, as defined here (i.e. the precision of scores), can be measured. This is usually done by generating more than one score for the same variable in each case, e.g. by asking the same question twice or by having two independent coders code the same evidence, and by assessing how much the resulting scores differ.

The reliability (or precision) of these scores is the average difference between the scores for each case. The smaller this difference, the higher is the precision (or reliability) of the instrument. This way of calculating measurement reliability can only be used for quantitative scores (i.e., ration and interval variables). For nominal and ordinal variables, reliability can be measured by calculating the proportion of cases in which the different measurements result in identical scores. This can be expressed as an inter-observer, inter-rater, or test–retest similarity rate.

Example: three operationalizations of "project success"

Financial success: When a valid procedure for measuring financial success of a project has been developed, its reliability can be assessed by arranging that two or more persons, either company staff, or researchers, or their assistants, collect and code data using these guidelines and then compute the degree of success from these data. The reliability of these scores can be expressed in terms of precision, i.e., as the average difference between the two (or more) scores. If the reliability of the scores is insufficient (in terms of a criterion that was formulated *a priori*) measurement procedures should be further specified until a sufficient level of reliability is achieved.

Delivery time: If a valid procedure for measuring the exact dates of planned and actual delivery and for determining its timeliness is developed, the reliability of the score can be assessed by arranging that two or more persons identify both the planned and the actual delivery date and then rate the delivery's timeliness. Assuming that there are only two possible scores ("on time" or "too late"), the reliability of these measurements will be expressed in terms of the proportion of cases in which the different raters have generated the same score.

Satisfaction: When a valid procedure for the measurement of the value of the company's project evaluation is developed, the reliability of the scores obtained in this way can be assessed by using the same procedures described above for assessing the reliability of financial success or timeliness of delivery. If evidence is extracted through qualitative interviews with persons, the more structured a qualitative interview is (e.g., instructions regarding the interview as well as the questions specified in the interview guide), the more reliable will be the evidence generated in the interview. Evidence generated in interviews by different interviewers with the same person should obtain similar or the same evidence. If the data are generated through a standardized questionnaire, consisting of questions with a set of response categories, reliability is usually assumed to be good, although different measurement conditions (e.g., how the questionnaire is introduced to the respondent, the absence or presence of other people such as supervisors or colleagues, whether scores are obtained in an interview or by self-completion) will influence the reliability of the scores that are obtained.

Back to our household examples: The score on a set of scales for body weight might vary an ounce when weighing the same body weight several times and we might find that "reliable" (precise enough), but for our kitchen scales we might want much more precision. Similarly, for outdoor temperature we might find a difference of one centigrade between readings too imprecise (unreliable), but one centigrade difference between two readings of the temperature in a kitchen oven might not bother us.

As mentioned above, using informants for measurement (by sending them a questionnaire) is a threat to measurement validity. It is also a threat to reliability, depending on how seriously the informant conducts the measurement (or tries to remember something).

Note that Cronbach's α is an attempt to assess measurement validity, not reliability.

Accuracy

If a measurement is valid and has a sufficient degree of reliability (precision), a third and final quality criterion is that the score is accurate. For instance, a set of scales can give a valid measurement of weight and the score on that scale can have an acceptable degree of reliability (precision), but it can consistently give a score that is two units too high or low. This might or might not be important. For instance, if we investigate the hypothesis that adhering to a specific diet might result in successful slimming, we might not bother that weight scores are consistently two kilograms higher than the true weight of the participants in our study, because this will not influence an effect size that is expressed in terms of the number of kilograms lost (assuming that we use the same set of scales for every measurement and accuracy of the scales stays the same during the study).

Evaluation of the quality of a study's measurements

The aim of question 6 of the checklist is to support a reader in assessing the quality of measurement in a study. It contains questions about the different quality criteria discussed above. These questions should be answered for every variable separately.

6. How are the IV and DV measured?
    a. Does measurement rely on informants or respondents?
    b. Are informants or respondents trustworthy?
    c. Are measurements valid?
    d. Are measurements reliable?
    e. Are measurements accurate?

A research report should allow a detailed assessment of these aspects of measurement by reporting for every variable how evidence is sought, observed, registered, and coded. The report should give information about how the researcher has evaluated the validity, reliability and accuracy of the measurements in the study.

CHAPTER 4
CRITICAL SYNTHESIS

This chapter discusses two main issues:

1. How is the result of a study discussed in a research report? (Question 7 of the checklist for critical evaluation of empirical evidence)
2. How can results of different studies be synthesized?

Conclusions from a single study

A theoretical claim (or "hypothesis") applies to a universe (or "domain") that usually is very large or even infinite, i.e., all consumers everywhere at all times, all firms everywhere at all times, etc. It is not possible to measure the hypothesized effect in this whole domain, because it is not possible (or at least not practical) to observe every single case in the domain. The best we can do is to investigate the effect in different subsets of this domain (which we call "populations"). Therefore, a theory-oriented research report (a "single study") is usually a study of just one population from the domain.

Results of a single study pertain only to the set of cases ("sample" or "data set") that is studied. If the sample is a probability sample from a defined population, then results can be inferred to that population, with a margin of error that is represented by the confidence interval. Because science is not a form of magic or revelation, no conclusion whatsoever can be drawn about cases (i.e., populations) that have not been studied. One of the extraordinary characteristics of published research reports is that in many of them it is actually claimed that general conclusions and managerial recommendations can be derived from the study. A research report can thus be evaluated in terms of the extent to which the author thinks that results can be magically generalized to other cases, other populations, other experiments, or even the future. Question 7a of the checklist for critical evaluation (*Does the report make claims beyond the studied population, i.e., a larger population; another population; the theoretical domain; the future?*) formulates this question.

Because results of a single study cannot be generalized to other populations than the one studied, the results of such a study can have only managerial relevance for the population that has been studied. If we take this limitation in mind, we can ask the question what exactly is the managerial relevance of the study's results for that population. In those cases, one can ask the question: If it is important for managers to change Y, is the effect size in the study large enough to make it worthwhile to engage in activities to change X? How much must X change in order to result in a relevant change of Y? This is question 7d of the checklist for critical evaluation (*What is the practical relevance of the observed effect(s)?*). Note that this question must be answered by an interpretation of the effect size (and its precision, expressed by a confidence interval), and cannot be answered by interpreting a *p*- or t-value.

In general, populations differ in important respects. Therefore, it is to be expected that effect sizes observed in different populations will differ. Hence, if the results of a study differ from the results of another study, one cannot draw the conclusion that only one of them can be right. (This mistake is made in current discussions in psychology, in which the failure to "replicate" a result is used as an argument for doubting the sincerity or professional skills of researchers who have published results that are different.) Obviously, results can be wrong, and every study must be critically evaluated, but it is only to be expected that results will differ between studies. The implicit assumption behind the expectation that results of different studies should be more or less the same is that the theoretical domain is homogeneous. This might be an acceptable assumption in natural science, and hence sometimes in medicine, but it is certainly not an acceptable assumption in social science as well as in business research. Homogeneity of the domain is the exception and should only be assumed *after* a series of studies in rather different populations have all shown more or less similar results. This issue is addressed by question 7b of the checklist for critical evaluation (*Is the theoretical domain assumed to be homogeneous or heterogeneous?*).

If a study's results are different from the results of other studies, it should never be assumed that this study is closer to having demonstrated a "real" effect size than any other study. Every result must be considered as information about the population that is studied and, assuming that the studies are good ones, differences between results should be interpreted as information about (real) differences between populations. The question then becomes how these differences can be explained. Hence, a quality criterion for the Discussion section of a research report is whether (and how) an attempt is made to compare and interpret the study's results with the results of other studies. This criterion is expressed in question 7c (*Are results compared (or synthesized) with those of other studies?*).

When all seven questions of the checklist for critical evaluation have been answered (in some detail), a conclusion can be drawn about different aspects of the quality of the study that has been evaluated.
If it is concluded that the results of a study cannot be trusted (e.g. if no probability sample was drawn in a cross-sectional study of a population; or if considerable non-response has occurred; or if measurement cannot be considered valid, reliable and accurate), then the study should be considered as providing *no empirical evidence* about the hypothesis. If the study was designed as a study of the hypothesis, then the study must be considered a *failure*.
If it is concluded that the results of a study can be trusted, but that the authors have not presented and interpreted these results in a correct or informative manner (e.g., by reporting "significance" and not reporting effect sizes), then one should make an attempt to "rescue" the valuable information that is contained in the study from the author's misinterpretations.

If it is our aim to generate a Critical Synthesis of the empirical evidence regarding a hypothesis, then we should select (only) those studies of the hypothesis that have trustworthy results and synthesize the results of these studies. It is quite common that we can only generate such as Synthesis after having found or computed relevant information about effect sizes ourselves, when that information was not reported in a study's report.

Critical Synthesis and meta-analysis

In theory-oriented research we want to find out how large the effect is of an independent variable on a dependent variable in different parts of a theoretical domain. We are interested in the variation of effect sizes through the domain: Is this variation large or small? What is the smallest effect size in the domain, and where do we find it? What is the largest effect size in the domain, and where do we find it? Can we find determinants of the variation or, in other words, can we explain or predict where the effect size will be large or small? Statistical meta-analysis can help us to find answers to such questions. The input of statistical meta-analysis consists of the effect sizes observed in the studies in different populations.

A finding in a single study can be close or far from the actual effect in a population due to sampling variation. Meta-analysis in its simplest form (using a "fixed effect model"; see Cumming, pp. 207-209) assumes that differences in findings between studies can be attributed entirely (or mainly) to sampling variation or, in other words, that all studies investigate the effect size in the *same population*. We take as an example for discussion a series of fictitious studies of response time to a word task, as presented by Cumming (pp.187-193).

Exercise 4.1

Take the **Original 31** page of **ESCI Meta-Analysis** and answer the following questions:

- What is the hypothesis that is investigated in this series of (fictitious) studies?
- What is the outcome of the meta-analysis?
- From which of the 16 single studies could one have predicted this meta-analytic outcome?

In this simple example the hypothesis is not about a relation between an independent and a dependent variable but only an expectation about the mean response time that will be observed in a study (namely 300 milliseconds). The mean response times in this example differ considerably between studies. For instance, in Study 13 (Mahov) a mean response time of 690 milliseconds is observed, whereas in Study 15 (Over) the mean response time is 314 milliseconds. The mean response time as calculated in the meta-analysis is just over 400 milliseconds; far more than 300 but far lower than 690. A first conclusion from this example is that *no conclusions can be drawn from a single study* of a theory. This shows that an effect size can only be *correctly estimated after a series of replications*.

Also note that the outcome of a null hypothesis significance test in a single study does not give any information that is relevant for an estimation of the overall or average effect size. If one would take the result of the study with the lowest *p*-value (here Study 13; highly significant, *p*-value not presented in Figure 7.5) as the most convincing result, as is often done, then one severely underestimates the "normal" or "average" speed of response. Note that Study 6 (Fox) has observed a mean response time that is very close to the average speed, but that the finding of this study in current routine practice would have been interpreted as giving no evidence for a longer response time than 300 milliseconds. This illustrates that *no conclusions can be drawn from the statistical significance of an outcome of a single study*.

Now consider a number of other issues regarding this example of meta-analysis (Exercise 4.2).

Exercise 4.2

Regarding the **Original 31** page of **ESCI Meta-Analysis,** answer the following questions:

- To what kind of people does the hypothesis (that the mean response time is 300 milliseconds) apply? In other words, what is the "theoretical domain"?
- What is the relation between the subjects in these 16 studies and the theoretical domain?
- How is it explained that the mean effects differ considerably between these 16 studies (from 314 to 690 milliseconds)?
- What is the main assumption that makes the meta-analysis as presented here?

The distribution of mean response times in this meta-analysis can entirely be explained as resulting from *sampling variation*. The assumption of a fixed-effect meta-analysis is that the distribution of findings represented in Figure 7.5 is caused by sampling variation only. Because the *true* effect size is assumed to be constant ("fixed"), different results from different parts of the domain can and should be interpreted as resulting from sampling variation (also called "sampling error") and measurement error, not as reflecting true differences. Statistical procedures of meta-analysis that use the fixed effect model "correct" for these types of "error" and estimate the true effect size that would have been observed if these errors would not exist.

But, consider the possibility that Mahov is one of the few researchers who does not conduct experiments with his own students (which is the routine practice in psychology, marketing, and most other disciplines) but instead with members of the "public". Would it still be defensible to assume that the difference in outcome between the Mahov study (Study 13) and the other studies is due to sampling variation only? And what if Golly (Study 7) and Over (Study 15) had conducted their studies with, e.g., trained air force personnel? Then the large differences in observed mean response times might be, at least to a large extent, be explained by real differences between the people studied and less by (still existing) sampling variation. Therefore, and more realistically, it must be assumed that the theoretical domain is heterogeneous and that differences in findings between studies must be attributed to actual differences in effect between the studied populations rather than to sampling variation. This is the reason why most researchers (and we in this book) use the so-called "random effects model" for meta-analysis (Cumming, p.209).

Because the effect sizes are expected to differ between populations, a Critical Synthesis of the extant empirical evidence regarding a hypothesis should not focus on calculating an "average" or overall effect size for the domain. Such an average has hardly any meaning. Much more interesting and relevant is generating (a) a description and quantification of the *variation* in effect size through the domain and (b) an understanding of the determinants of higher or lower effect sizes. "Simple Meta-Analysis" (SMA), a simple software tool for meta-analysis developed at the Rotterdam School of Management, generates a "prediction interval" of effect sizes in the domain, which is a description and quantification of the variation in observed effect sizes. It does not contain a possibility to investigate determinants of that variation. Other, more complex software tools for meta-analysis do contain such features.

REFERENCES

Abhijit V. Banerjee & Esther Duflo (2009), "The Experimental Approach to Development Economics", *Annual Review of Economics*, Annual Reviews, vol. 1(1): 151-178.

Denny Borsboom, Gideon J. Mellenberg, & Jaap van Heerden (2004), "The concept of validity", *Psychological Review*, 111(4):1061–1071.

Geoff Cumming (2012), *Understanding the new statistics. Effect sizes, confidence intervals, and meta-analysis*, New York: Routledge.

Jan Dul & Tony Hak (2008), *Case study methodology in business research,* Oxford: Butterworth-Heinemann.

Joseph Henrich, Steven J. Heine & Aya Norenzayan (2010), "Most people are not WEIRD", *Nature*, 466:29.

Diana C. Mutz (2011), *Population-based survey experiments*, Princeton University Press.

Steven G. Rogelberg & Jeffrey M. Stanton (2007), "Introduction: understanding and dealing with organizational survey nonresponse", *Organizational Research Methods*, 10, 195-209.