Chapter 2

ELS research methodology: general issues

2.1 Introduction

This chapter discusses a number of issues that are at the basis of empirical legal research methodology. They deal with preparatory issues such as research question formulation, research ideas and specifics such as hypothesis formulation. Extensive attention will be given in this chapter to quality aspects of measurement, such as validity and reliability of the measurement of constructs. We will do so for quantitative as well as qualitative research, and discuss reasons why combining the two – in a so-called mixed methods design – generally strengthens our conclusions. Next we discuss a number of commonly used categorizations of research questions and designs, at the micro- meso- and macro-level, and briefly touch on the distinction between primary and secondary data. We end the chapter with research ethics, the rules that researchers should stick to by law and disciplinary convention.

2.2 Empirical cycle

Researchers can come across a research idea at any time, in any place. No rules describe how researchers should identify a research idea. Some researchers come upon ideas when they are peeling potatoes, others while sitting in their study bent over thick books, others while jogging along a canal in the rain. No rules prescribe how you as a researcher should arrive at a new research idea.

Some refer to this phase as the 'observational phase': you as a researcher observe phenomena, you walk around and take notice of things – out of which an idea bubbles up. Obviously, quite often researchers are approached by governmental bodies who ask them to answer a certain policy- or practice-relevant research question.

This research idea is what starts the so-called *empirical cycle*. As said, this idea can be a very rough notion of what you as a researcher want to zoom in upon. However,

once you have identified a research idea (say: the relation between tort victimization and well-being, or the recruitment process of law firms) there are a number of steps to follow for that idea to be turned into a researchable question and for us as researchers to find an answer to our question. That sequence of steps is often referred to as the 'empirical cycle'. Qualitative and quantitative research may differ in how they progress through that cycle, and we start by describing the steps in the research cycle for quantitative studies.

So: let us suppose that the research idea has been identified. After a research idea has been decided upon, it should next be translated into a question. This phase is what is called the *inductive phase*. Here the rough notion is transformed to a question to which you want to find the answer. That question could be, for our example: are criminal law judges less intelligent than civil law judges? (It could also be: is there a difference in intelligence between criminal law judges and civil law judges?) These two varieties may seem only superficially different but, as we will see in chapter 9, the exact formulation of the research question does have implications for statistical testing. This inductive phase is also referred to as the phase where hypotheses are formulated, or the 'hypothesis phase'.

Suppose that the research question we choose is: 'are criminal law judges less intelligent than civil law judges?' Then, before we can start carrying out our research to provide an answer to this question, we must become much more concrete. It is for instance unclear in this question what we mean by intelligence. Intelligence is what is often referred to as a 'construct' or 'hypothetical construct': something people have a common understanding of, but something we cannot observe directly like we could someone's height or weight or eye colour. We assume that it exists, but it is not observable, it is only hypothesized to exist. As the construct is not an observable property, we have to define how we measure the construct, in this case: intelligence. Do we intuitively infer it from someone's appearance? Do we infer it from someone's educational level? Or do we use someone's score on a validated IQ test?

Next we have to decide what we mean by a criminal and civil law judge. Is that someone who has worked as a judge only in criminal or civil courts respectively? Or someone who has done so over the past five years, regardless of what other type of cases s/he tried? Or who is doing so currently? And how do we measure this? Do we ask judges in what type of court they worked? And whether they did so exclusively? What do we do with those who have experience in administrative cases as well? Suppose we do our study in the Netherlands, do we include judges at the International Criminal Court?

Subsequently, we also need to decide when we will answer our question in the affirmative. Suppose that criminal judges' average IQ is 120.0000, and that that of civil law judges 120.0001, do we then answer our question in the affirmative? Or do we only identify a difference as 'telling' from a certain magnitude onwards (for instance a 10-point difference, or a 5-point difference)?

In this phase the research question or hypothesis becomes testable. It is specified to such an extent that we will actually be able to find the answer to the questions through measurement. From the general question that was formulated in the inductive phase in which we wrapped the question in vague words, we have now come 'down to earth' in this next phase, the *deductive phase*, and defined exactly how we are going to *measure*

the concepts or constructs. It is sometimes jokingly said that after the deductive phase, after the questions have been turned into testable hypotheses, if the researcher were to suddenly be run over by a tram in front of the university building, the research could go ahead as planned. After his or her untimely demise another researcher could take over and would conduct the research in exactly the same way as the deceased researcher would have done. All decisions have been made, all steps have been outlined. The path or 'recipe' for the study has been outlined. This process, of defining exactly what is meant by a construct and how it can be measured, is called *operationalization*. Operationalization is the process through which general ideas are turned into concrete to-do lists. With operationalization, you lay down exactly through what operations you are going to answer the research question. After operationalization, the research has become *reproducible*, by which we mean that if asked to re-do the study, the authors would be able to do so.

Reproducibility – the extent to which the original researchers can reproduce their findings – is different from replicability, the desirable situation where findings conducted in one study by a set of researchers are replicated in a different study by different researchers (it should be noted that replicability and reproducibility are at times defined differently in different disciplines). Replicability can also be regarded as a form of external validity, namely the situation where the conclusions from one study generalize to other contexts (see section 2.6.2).

After the deductive phase, the research proper can go ahead. According to the precise prescriptions, measurements can be made and we will be able to determine whether civil law judges are indeed more (or less) intelligent than criminal law judges. This phase is often referred to as the 'testing phase', that is, the phase where data are collected in the empirical world, to test the hypothesis.

Lastly, after the hypothesis has been accepted or rejected, the last phase encompasses the consideration of this result and its meaning for theory. Suppose – for our silly example – that criminal law judges are found to be on average indeed less intelligent than civil law judges, what are the implications for our empirical legal theories? Can this finding be due to an effect in the sense that the cleverer law students are more often assigned to the 'superior' civil bench? Or is it that after working for many years as a criminal court judge, your mind becomes numb and your IQ decreases? Or is there perhaps a *confounder* at play, in the sense that the criminal law judges are older (it is known that IQ decreases with age)? Here, in this so-called *evaluation phase*, we are assessing what the findings mean: is it really true that the type of chamber in which a judge works is associated with intelligence. And suppose that were the case, of course we would immediately ask ourselves why that would be? Clearly, an answer to the first question immediately raises a follow-up question to be studied.

In that sense, we generally have not come full circle after these five steps. And indeed, the empirical cycle is generally not regarded as a cycle, but as a spiral, with each answer to a question generating new research ideas. Viewing the empirical cycle as a spiral rather than a cycle concurs with the idea that research is accumulative and leads to knowledge in small steps, with each step and each subsequent study constituting progress and taking us further along the path to understanding the world around us.

The research cycle as we have described it just now is characteristic of positivist, quantitative research. It is much less typical for qualitative research. In qualitative

research, testing is generally not done, nor are 'instruments' (in the sense of an instrument such as a questionnaire to measure intelligence) commonly used. This is mainly due to the fact that the types of questions posed in qualitative research are generally different. In qualitative studies, we are generally not so much interested in the magnitude, frequency or scale of phenomena. Rather, we ask questions referring to the 'why' or the 'how' of phenomena. For instance, we may be interested to know why people like to work as a barrister. Obviously, we could also answer that question using quantitative research. We could present members of the bar with a set of answers to this question (for instance: a. 'Because I want fast money'; b. 'Because I want to change the world'; c. 'Because I know nothing better to choose'), but that is not the manner in which qualitative studies aim to answer questions. Rather, qualitative studies aim to always study phenomena within their context. Thus, here, a researcher would not simply sit down with a number of barristers or e-mail them questionnaires, but would – if s/he chose a more ethnographic approach – immerse him- or herself in the bar, in staff meetings, sit in on cases being tried, talk to the lawyers also after work, drink beer with them, bike with them to the day care centre where they pick up their children after work. As such, the researcher would try to *understand* through frequent interaction, observation, getting to know barristers, why decisions such as applying for the bar are made, and what gratifications its member receive for their work. The understanding of the subject matter would be deeper, contextual and in all likelihood more variegated. Researchers might also stumble upon motivations they could not have foreseen. Qualitative research might be able to uncover how for different people, a different set of motives operated, depending on the particular phase in their life and choices they were facing, or depending on child care support they could enlist, or not. For that matter, qualitative research is also referred to as *interpretative* empirical, with quantitative research conversely labelled as *analytical* empirical.

The way qualitative research operates is therefore generally much less prescribed and much more open than quantitative research. At the most qualitative end of the spectrum, researchers often start by interacting with the people they aim to study, and draw up hypotheses as they go along. In such studies, each interaction or interview may be seen as a separate mini-research cycle, after which each time new research questions or explanations (not often referred to as such, though) are distilled, which are then tested again with every new research activity. A qualitative study is in that sense by itself a spiral of small research cycles. By the time the explanation or interpretation of the phenomenon under study converges, the research ends. The researcher then has an overview of the phenomenon.

In qualitative research, the research cycle can even be inverted, with researchers starting with data collection, and formulating questions on the basis of what they observe. This type of research is referred to as *grounded theory* and the data gathered in this fashion as *grounded data*; the 'grounded' refers to the fact that the questions thus formulated are based on empirical observation, and not decided upon before observation starts. As Strauss & Corbin have said (1990, p. 23), a theory is:

> "*inductively derived from the study of the phenomenon it represents. That is, it is discovered, developed and provisionally verified through systematic data collection and analysis of data pertaining to that phenomenon.*

> *Therefore, data collection, analysis and theory stand in reciprocal relationship with each other."*

Grounded theory (see Glaser & Strauss, 1967 and Strauss & Corbin, 1998 for the classic references, and Charmaz, 2014 for a more recent update) is an approach that is useful or even necessary if one wants to study phenomena that are largely uncharted terrain. As Stern (1995) has said:

> "*the strongest case for the use of grounded theory is in investigations of relatively uncharted water, or to gain a fresh perspective in a familiar situation."*

We have so far described qualitative and quantitative research in a fairly caricatural manner: qualitative research as open and grounded, with researchers immersing themselves in uncharted communities, and not working from any existing theoretical notions or categorizations, and quantitative research on the other hand as inflexible and deterministic and without any options for open answers or contextual understanding. In practice, the twain regularly meet. Quantitative questionnaires however often also contain some open-ended questions, and qualitative researchers may also quantify their data to be able to tally or compute correlations. They had perhaps, for that reason, best not be seen as entirely different unmarriageable approaches but as complementary, with mixtures and blends being quite feasible.

2.3 Literature search and source evaluation

Often the research idea, the first phase of the empirical cycle, is not a sudden revelation or a lightning bolt. In fact the idea phase often coincides with the period where reading is done. One may for instance be reading up on the scientific literature on employment careers, and find a number of studies that postulate that many lawyers end up at the bar because of the networks they built when they were students. Such 'pointers' from previous research often generate new research ideas.

In general, it is in fact highly inadvisable to start out on research without a thorough literature search. Once a vague idea or direction for the research has been formed (such as 'I want to know more about barristers'), a first exploration should be done of the scientific literature. Reading recent work on the topic tells you what has already been found, what is still an enigma, what new directions are considered fruitful by the cognoscenti in the discipline. Reading up will also warn you against embarking upon topics that are very hard to study in practice, and reading about other researchers' struggles will save you time and money, and prevent you from making the same mistakes.

Literature searches can be done through the internet, or in literature databases. Not all academic literature is 'open access' yet (meaning it can be read without charge) and for many articles in scientific journals you still need to search their content through a university e-mail account that grants you access. It is important to be (extremely) critical when using information from the internet. For scientific purposes, it is in fact best to use only sources that have been published in academic, peer-reviewed journals. Even though that does not guarantee high-quality research, it does give a layer of

checking that was done by journal editors and reviewers. Academic work published in academic peer-reviewed journals is ranked as more noteworthy and telling than studies published in non-academic or non-peer-reviewed journals, or work published on researchers' own websites. Top-ranking journals are *The Lancet* (for medical research), and *Nature* and *Science* for various disciplines, mostly from the harder sciences – the latter two so highly ranked that they are in a sense in a separate universe.

For empirical legal studies, a few journals stand out. The first of these is the *Journal of Empirical Legal Studies* (`onlinelibrary.wiley.com/journal/17401461`), with a relatively select and small readership. A second journal that publishes empirical work (with a socio-legal focus) is the *Journal for Law and Society* (`onlinelibrary.wiley.com/journal/14676478`). The European Society for Empirical Legal Studies (`esels.eu`) aims to publish its own *European Journal for Empirical Legal Studies* as of 2024.

Searches are generally done using keywords such as *lawyer*, *barrister* or *legal professional*. In describing the findings, it is common to list, apart from the keywords, the databases searched, the date the search was conducted and the number of hits. Your output is then reproducible. Many researchers next 'snowball' in the sense that they scan the reference list of each of the publications found, retrieve those publications, in turn scan the latter's reference lists, and so forth.

Some academic conferences publish abstracts on their websites that can also be searched. Often, these do not contain sufficient material to refer substantively to these conference presentations, but it is not uncommon to e-mail the presenters and ask them for their PowerPoint slides or any unpublished papers they may have.

2.4 Conceptual and operational definitions

If we want to make statements about the explanatory power of theories, we need data. Empirical data are necessary to test for instance whether the extent to which suspects' experience of procedural justice accurately predicts to what extent they were content with the sentence handed down. We could sit in and witness a large number of randomly selected criminal law cases, note to what extent the defendant was allowed to speak, whether s/he was addressed by the judge in a respectful manner, and in general tick off all the other variables that are assumed to be relevant in the theory of procedural justice (Tyler, 1990). Next, after the sentence has been handed down, we interview the defendant and note his or her 'contentedness' on some scale with the outcome. If we find correlations between the various aspects of procedural justice and the extent to which defendants were content with the outcome, this supports the theory of procedural justice. If we find no such correlation, we have little empirical support for the theory.

If the theory says that – regardless of the severity of the sanction – defendants who are addressed in a condescending manner will be less content with the outcome than those who were treated with respect, and if the theory holds, we should have found the patterns as predicted by the theory. But if that is not the case, something is wrong. Did we investigate the type of cases that the theory supposedly holds for? Should the theory be adapted? If measurements fail to support theories, this can also be because

the operationalization of the variables was different from what the theory purports they should be.

Many theories employ so-called hypothetical constructs, abstract notions that in quantitative research need to be defined precisely before they can be measured. Earlier, we used 'intelligence' as an example of such a construct.

Tyler (1990) is regarded as having minted the construct of *procedural justice* (also encountered as 'procedural fairness'), by which it is meant that legal procedures are seen to be conducted in line with principles of fairness. Procedural justice is assumed to be an important pillar of legitimacy. The theory of procedural justice posits that if citizens regard the justice process as having been conducted in line with fairness principles, they are more likely to comply with the outcome, even when the outcome of the process is unfavourable for them. Formulated differently: the theory posits that how citizens regard the justice system is tied more to the perceived fairness of the justice process (including the manner in which citizens are approached) than to the perceived fairness of the outcome.

The construct of procedural justice is generally regarded as being multidimensional, although these dimensions are encountered in the literature in slightly different constellations. Notable dimensions are (1) voice (citizens are given the opportunity to express their side of the story); (2) respect (officials treat parties with dignity and respect); (3) neutrality (the decision making process is unbiased); and (4) transparency (parties are able to see the above being done). Other dimensions that may be postulated are (5) understanding (citizens understand the process and how decisions are made); and (6) helpfulness (perception that system players are interested in your personal situation to the extent that the law allows).

This shows that, while scholars may employ one and the same wording for a construct, their exact definition of that construct may regularly vary. Also, definitions of the same term may evolve over time. A good example again is the hypothetical construct 'intelligence'. While the word intelligence is common usage in many languages, numerous definitions and measurement strategies have emerged over the years it has been employed in psychological research. In the early years, the Stanford–Binet test was used to measure people's intelligence level. As the science of psychology evolved, more dimensions were added to the construct, particularly capturing nonverbal abilities. Currently, many tests are available, such as the WISC or the Raven, which each operationalize intelligence (slightly) differently.

After a construct has been defined, after we have described what exactly we understand the construct to be, through a *conceptual definition*, we next need to decide how the construct should be measured. This is done though a so-called *operational definition*. If for instance our conceptual definition of intelligence states that intelligence is the ability that individuals have for language, maths, and logical reasoning, our operational definition would likely state that our test must contain items to capture language skills, mathematical skills, and logical reasoning. If our conceptual definition also includes visual-spatial processing skills, then the operational definition would state that the test also contains items for these.

All this applies strongly to quantitative research. As said above, the definition of constructs in qualitative research is generally done much more inductively, working from empirical reality with constructs emerging from it.

2.5 Pilot study

In general, once the literature has been read, and the variables have been operationalized, studies do not go full steam ahead yet. It is in fact recommended to test all the study's elements before the study goes 'live'. The study's manner of contacting the respondents should be tested, any instruments should be tested, questionnaires should be administered to a number of trial respondents to see whether filling out the forms really does not take too long, whether all the questions are understandable to respondents, non-offensive, and the like. Even interviewers should be 'tested', as well as research protocols, apps that are used, ICT settings. Such a general rehearsal for the study is referred to as a 'pilot study'.

Studies should be piloted before they start. In spite of all the good intentions and years of experience of research staff, it turns out that – as the saying goes – in research too the proof of the pudding is in the eating. Only once the research is actually carried out in the field, can we find out whether it works. And often that research pudding turns out to need a few more tweaks before it is really up to standard. Thus, almost all researchers pre-test the crucial elements of their studies, and almost all find they need to adjust the study at least a bit. Not piloting one's study is considered extremely risky.

Piloting is especially important for quantitative research. Once a questionnaire has been printed 2,500 times, or distributed to an e-mail list of 5,000 respondents, changes cannot be made without great damage to the study, and concomitant costs. That is why it is so extremely important to have a good test-run before the research goes live. In qualitative research, the need for testing is much less crucial. Of course, some kind of testing is inherently necessary, especially with regard to obtaining access to the people one wants to study. Much of qualitative research is however flexible and adaptable, so that as the research proceeds, changes to questions can be made, questions added, the protocol adapted, etc.

2.6 Validity and reliability

We discussed above the need to operationalize constructs such as intelligence. But how can we be sure that our operational definition does give us good measurements of the construct we intend to measure? Perhaps the items with which we aim to measure intelligence have been chosen such that it is easy to guess the right answers. Or perhaps maths skills items have been formulated so 'wordily' that they actually tap language skills, so that only those with sufficient language skills are able to understand what is asked – in that case the test does not capture maths skills.

This issue – whether our operational definition enables us to measure the construct as we defined it – is generally captured in two methodological properties: validity and reliability. Let us first turn to reliability.

2.6.1 Reliability

Reliability refers to the exactness with which we measure our construct. Suppose I would want to measure someone's height. I could do so using a measuring tape. If my

measure of height is reliable, this means it is exact: each time I measure that person's height, I get exactly the same result. In the reverse case, namely if I get different results each time I take my measurement, I call my measure unreliable. Using an elastic measuring tape would for instance make for an unreliable measure: one time the result is 170 cm, next it is 172, then it is 167, then 173 etc. If I were to do my measurements 100 times I might on average get the right result, but because of the variability in the measurements, we say that the measure is unreliable.

How can we assess a measure's reliability? The example already gives an indication of that: one way to do this is to test and then re-test: if we obtain the same results upon re-testing we call the measures reliable. If we do not, we call them unreliable. Testing and re-testing is not always an option, however: if we for instance interview respondents, it is not feasible to return to respondents and re-interview them. Even if the objects we are investigating would lend themselves to re-testing (for instance when we are conducting a dossier analysis), it still often does not make sense to re-test: the coder would probably remember the value s/he entered previously.

In such cases, we tend to have two coders investigate the material, independently, and compare their measurements. If the results from these two raters concur, we say that we have *interrater reliability*. Reliability is often expressed as percentage agreement: it is then reported, for instance, that for the variable that measures the number of children over whom custody issues arise in divorce proceedings, interrater agreement is 90, meaning that in 90% of cases the coders arrived at the same value for the variable. Percentage agreement varies from 0 to 100.

Percentage agreement is often regarded as a fairly simplistic measure of interrater agreement. It namely disregards the fact that some agreement may occur by chance. A better measure, which is often preferred, is Cohen's kappa (also written with the Greek letter κ), a statistic that takes the observed level of agreement between coders and corrects for agreement that would have occurred by chance. Kappa ranges from -1 to +1, with 1 indicating perfect agreement, 0 indicating as much agreement as expected by chance, and -1 indicating perfect disagreement. There are no hard rules to judge whether interrater agreement (which will seldom be perfect) is good enough. Most authors rate agreement in a stepwise fashion, with for instance absolute values of 0.41 to 0.60 indicating moderate agreement, 0.61 to 0.80 indicating sufficient agreement, and 0.81 to 1.0 indicating almost perfect or perfect agreement (Landis & Koch, 1977). Others are more strict and believe that only kappa-values indicating that two-thirds or more of the rankings are identical are acceptable.

But both percentage agreement and Cohen's kappa may be less than optimal measures of reliability if we have a numerical variable (such as number of children). Percentage agreement and Cohen's kappa simply look at whether two coders fill out the same value, but they disregard whether the difference between scorers is small ('4' versus '5') or large ('1 child' versus '5 children'). One solution that may be employed is to use a measure of association, such as the correlation coefficient (see section 6.4.1). Versions of kappa are also available that weight the differences between coders. In case the scorings of more than two raters are to be compared, a special statistic should be computed to arrive at an overall measure of consistency across raters (an example is the so-called intraclass correlation coefficient or ICC). For a fairly technical but instructive and comprehensive overview see Hallgren (2012).

Reliability may be less easy to assess in qualitative research. As the qualitative method does not often rely on instruments, and as much of the data may be gathered in interaction with the research subjects, it may be the case that – as is often said – it is the researcher him- or herself who 'is' the instrument. As interactions can only occur when at least two people are present, interactions could not even be replicated through a second researcher or coder.

However, when text has been coded or video material, we have data that is 'patient' and that can be re-coded by another researcher. Then, it is possible to assess reliability. The two (or more) coders then code fragments of text, and one can check whether they attach the same labels to the same chunks of text. Krippendorff's α is a coefficient of agreement between coders that is regularly used when textual units have been coded. It was developed for content analysis, which we will discuss in section 7.3. When Krippendorff's α is 1 there is perfect agreement; when it is 0, agreement is no higher than would have been achieved by chance. It can be used with any number of coders. For more, see Krippendorff (2004, 2011). We return to the topic of reliability in qualitative research in section 7.5.

2.6.2 Validity

Next, validity refers to the extent to which our instruments really measure what we are after. Where reliability has to do with the extent to which we measure our construct precisely, meaning that it would be replicable, validity is much more fundamental in that it reflects to what extent an instrument to measure, say, criminogenic beliefs, actually captures those beliefs and not something else. Designing instruments such that they truly reflect the construct you are after is a skill, and requires extensive testing and probing before you can be confident that the instrument is up to standard.

Focusing again on quantitative research, let us illustrate with an example some of the issues that complicate achieving valid measures. Suppose I would indeed want to measure criminogenic beliefs. Some of the questions I could pose respondents are, for instance: 'It is okay to steal', or 'If someone does not pay me respect, I am willing to use force'. Now some of your respondents who harbour criminogenic beliefs may answer 'yes' to the first question, and 'yes' to the second one. Most respondents with criminogenic beliefs will, however, realize that their beliefs are considered 'not done'. Anticipating disapproval, they may therefore choose to answer 'no' to your questions and present themselves as respectable citizens, just like they believe the interviewer to be. If that happens, we say that the instrument is sensitive to 'social desirability bias'. The instrument then does not capture criminogenic beliefs, but to a large extent what the interviewee believes the interviewer finds a 'correct' answer. Your instrument is then not valid. More tendencies exist that influence respondents' answers and thus affect validity, and we will discuss some additional ones in chapter 4.3.3.

In quantitative social science research methodology, several kinds of validity are distinguished. The first one is the one that we just referred to: does the instrument measure what it is intended to measure? This is denoted more specifically as *construct validity*. For an instrument to have construct validity, it has to meet three criteria. First, the instrument must capture the construct across its entire 'spectrum'. For instance, an intelligence test that has maths items only does not capture intelligence as we under-

stand it: it does not measure language skills or logical reasoning or nonverbal abilities. It therefore assesses the construct only partially. It is then said that the instrument lacks *content validity*. Only an instrument that assesses the construct across its entire breadth or spectrum, is said to have content validity. Second, the instrument must generate measures that correlate in the expected direction with other outcomes. This likely sounds fairly obscure, so we will try to clarify it with an example: the scores on an instrument to measure driving skills should be associated negatively with the likelihood that people cause traffic accidents. If the instrument tells us that someone has good driving skills, we would expect that person to drive safely. This is referred to as *criterion validity*: the scores obtained with the instrument should correlate with an external criterion that you would expect it to be associated with. Third, we want our instrument to really capture the underlying construct, and not something else. It is for instance possible that we might design an instrument that has good content validity (say we are again measuring intelligence, and our instrument encompasses intelligence across all its dimensions), and also has good criterion validity in the sense that people with high scores on the instrument do well in school. However, if our instrument is built using items written in advanced Dutch, it does not capture intelligence but mainly language skills. It would for instance be unable to measure intelligence for recently arrived migrants who have not yet mastered Dutch. Then this instrument still does not measure intelligence. This last criterion is referred to as *construct validity-in-the-narrow-sense*.

All in all, an instrument can only be said to have validity if it meets all of the three criteria: content validity, criterion validity and construct validity-in-the-narrow-sense. If all these three criteria are met, can we then say that the instrument has construct validity? Not yet, as for an instrument to have construct validity, we also require it to be reliable. Formulated in ordinary terms: we cannot say that our instrument measures what we are after if it produces imprecise measurements. So, in summary, we regard an instrument to have construct validity only when the instrument gives us precise measurements of the construct we are after *and* meets all of the three above-described validity criteria. The instrument then has full construct validity.

How then does one assess whether one's instrument has construct validity? This is a lot more complicated than assessing reliability. Researchers generally first assess *face validity*, which may sound sophisticated but is not much more than judging whether, on the face of it, the instrument is likely to be content, criterion and construct (in the narrow sense) valid. More intricate methods investigate whether the scores generated by the construct correlate with other variables in the way they should if they were a valid measure of the underlying construct.

Other kinds of validity exist. We discuss them briefly. *Statistical conclusion validity* or *statistical validity* refers to whether a result is likely to be attributable to anything other than chance. Suppose that we have investigated 100 volunteers, young men aged 15–25 years old. We have administered our intelligence test and asked them whether they have ever been a victim of an accident. Now suppose that we find that our measurements of intelligence correlate with our respondents' accident victimization: the lower the respondents' intelligence scores, the more accidents they suffered. Statistical tests can tell us how likely it is that the observed association is a chance one. If our statistical test tells us that the observed correlation is in fact a highly unlikely outcome if intelligence and accident proneness were unrelated, we say that the correlation is

significant. We then conclude that intelligence and accident proneness are related. The finding then has 'statistical conclusion validity'. We discuss statistical testing in detail in chapter 8.

Next, *internal validity* is distinguished. Continuing our example on intelligence and victimization, we might be tempted to believe that, given the statistically significant association between the two, intelligence and accident proneness are in fact causally related, with people with lower intelligence more often suffering accidents: because they do not pay attention, do not learn from past mistakes, etc. We cannot be sure about this, however. It may for instance be that intelligence has nothing to do with accident proneness, but that those with higher intelligence do not like to own up to the fact that they were involved in accidents (in which case our instrument could be said to suffer from social desirability bias.) Or could the lower intelligence score be due to the accident suffered? Or could it be the case that people with lower intelligence work in different jobs that expose them to more risk of accidents?

As there are several alternative explanations – other than that low intelligence causes accidents – for our finding, we do not have internal validity. Even though we captured intelligence with a construct valid measure and even if we have valid measures of the number of accidents that people suffered, and even if the association is significant, we still cannot rule out other explanations than the causal one. This happens very often in observational research: researchers find associations between an intervention and an outcome, but they cannot prove that the outcome has been caused by the intervention. If vegetarians live longer than meat eaters, this may also be due to the fact that vegetarians do more exercise, or smoke less, or are generally more health-conscious. If marathon runners live longer than non-runners, this may also be due to the fact that they are healthier in the first place (which enables them to do those gruelling runs and makes them live longer, so that it is actually a reverse causal relation). These other variables, such as exercise, smoking or health (consciousness), are called *confounders*.

Ensuring internal validity is extremely difficult in observational studies like the examples given here, and in fact in much empirical legal research. The only way to be absolutely certain that a causal relation exists between for instance participation in an anti-discrimination training for municipal workers and the outcome of citizen satisfaction is if we could be certain that those who participate in the training do not differ from non-participants. The only way to be entirely sure about this is to administer the intervention at random: just like in pharmaceutical research, we should then create two groups who randomly do and do not get the anti-discrimination training. If we have such a design, we will have ruled out any alternative explanation. As the two groups (those with and those without the intervention) have been formed by chance, there are no systematic differences between the two groups anymore and they therefore also cannot differ anymore on the confounders. Then, no confounders can be at play, and no alternative explanations are left – only the causal one. We will return to this topic at length in chapter 5.

Lastly, suppose that we have a study conducted with a measure that has construct validity, and suppose our findings have statistical validity, and suppose as well that we have internal validity: can we be certain that trust in the municipal institutions as measured in our sample would also be found in the entire population once all municipal workers had been trained?

External validity can also be conceived of in a broader sense. Suppose that we have studied a pilot neighbourhood court in Rotterdam, and that we found that the court is able to effectively and speedily deal with the underlying problems of defendants (addiction, homelessness, debts, etc.), so that processual costs are massively reduced. We can ask ourselves the question whether we may be confident that we would find similarly successful outcomes if neighbourhood courts were to be rolled out in other cities? More generally: can we be sure that interventions tested in certain 'experimental' settings such as research labs or pilot projects would also be found in the outside world, outside of the artificial research setting induced by academia or the enthusiastic innovative pilot? If that is the case, we have *external validity*: our findings are then also generalizable to other situations than our artificial research environment. We will return to the issue of generalizability from samples in section 3.2.

Note that the four kinds of validity we just discussed constitute in a sense levels or steps: each validity level has to be satisfied before it makes sense to check whether the next type of validity is present. If we do not measure what we are interested in, it does not make sense to test whether a result is statistically significant. If an association is not statistically significant (and is therefore likely to be chance result), it does not make sense to investigate whether an association is causal. Therefore, with each successive validity level reached, we become more confident that our findings are meaningful.

All that has been discussed so far on validity pertains particularly strongly to quantitative research. Assessing validity in qualitative research – especially construct validity and internal and external validity – is often done slightly differently. We return to the topic of validity in qualitative research in section 7.5.

2.7 Qualitative and quantitative studies

An often-used categorization of research, which we have already briefly touched upon, is that into qualitative and quantitative studies. While quantitative studies aim to measure the volume or *quantity* of some variable of interest, qualitative studies are geared to find out the 'why' or the 'how' of phenomena. Quantitative studies typically follow the empirical cycle, are studies where hypotheses are formulated and where statistical testing is employed. For understanding, for instance, why corporate fines do not reduce environmental crime, or what role family lawyers play in the deflection of divorce conflict, qualitative studies may be much better suited.

When we conduct qualitative studies, much is different from what has been discussed up to now. Qualitative studies are generally much less strictly formatted, use lots of unstructured interviews and observation, have more room for adaptation as the research goes along, and leave more room for respondents to put forward unexpected viewpoints. They may be ethnographic, in which case the researcher spends time on a structural basis with his or her research subjects, to understand their world and reality, and to be able to explain 'from within'. Instruments are hardly ever used, statistical testing is hardly ever encountered.

While quantitative research gives a broad, generalizable, quantitative summary of a phenomenon (which are plus points), its downside is that it is reductionist and can deal with only a small number of factors in relatively simple models (which are neg-

ative points). Qualitative research is richer in the sense that it can describe complex situations, uncover mechanisms, understand behaviour in context and deal with multiple layers of interrelated causal factors (these are all plus points), but often has limited generalizability and limited reliability (negative points).

In qualitative research, the aim is much less to produce generalizable quantitative statements, but rather to unravel a number of mechanisms, to *understand* what happened, within a certain context. Some qualitative researchers even may find lack of generalizability unproblematic as they regard all behaviour as contextual. This is why the inherent juxtaposition of qualitative and qualitative research is often portrayed as we did in Figure 1.1 in chapter 1.

A quantitative study gives in general a broadly generalizable result. It has nice properties and strengths, amongst which its generalizability, and the fact that all the data that are collected and analyses that are carried out are easily replicable. Other researchers can use the same instruments, draw a sample in the same manner from the same population and would find (approximately, as there will be some chance variation) the same result. In that sense, quantitative studies are strong on reliability, reproducibility and replicability. However, quantitative research is generally limited in the kind of models it can analyse: only so many variables can practically be analysed simultaneously. In addition, quantitative studies can only give you what you were searching for: the questionnaires will generate answers only to the questions that were posed, not to other useful things respondents knew but had no space in the questionnaire to tell you about.

A qualitative study, on the other hand, focuses generally on one or a few particular settings, and investigates behaviour in the complex environment in which behaviour emerges: in interaction with others, understanding and misunderstanding cues, expecting things and asking for one thing but wanting another, analysing symbolic meanings that people give to behaviour, dress code and the like. Qualitative studies can be serendipitous. This makes qualitative studies attractive for getting to the heart of things, for gathering new insights. Phrased differently, qualitative studies can be stronger on validity. The downside is that it will be very hard, if not impossible, to replicate and sometimes even reproduce qualitative studies; the researcher him- or herself may in some cases even be regarded as the 'instrument' in a qualitative study. It is through the interaction with the researcher, through the impact that the researcher had on the respondents, through his or her observations and interpretations that the analysis takes shape.

This may all sound pretty abstract as we have not given many illustrations of what qualitative and quantitative studies are about. We will do so in a stepwise fashion in the next chapters. For now, it is important to remember that a quantitative study is generally relatively strong on reliability, a qualitative study relatively strong on 'validity' in the broad sense of unravelling mechanisms, capturing the intricacies of processes – all, obviously, given a high quality of the study design, measures and staff.

2.7.1 Mixed methods and triangulation

While the two types of research traditions are at times juxtaposed, and researchers from the two traditions often move in their own separate circles, this writer believes

that qualitative and quantitative methods can very well be combined. Contrasting the two is an unnecessary 'antagonistic' schematization as, first, no type of tradition has a priori preference over the other. Suppose we are interested to study the 'how much', or 'change' in a variable, for an entire population. Then a quantitative study is clearly best used. If the question on the other hand is one about the why or the how, or a question on a topic that we know very little about, then a qualitative study is likely best chosen.

Second, instead of seeing the two traditions as different, they are perhaps better regarded as complementary. In many instances we are namely not only interested in the – putting it in a very black-and-white way – 'how much', but for sure also in the 'how' or 'why'. And even if we are interested in barristers' motivations for working at the bar, we might still want to have some idea of how often certain motivations prevail, or whether motivations differ for younger and older respondents, or men and women. It is in practice in fact often the case that researchers combine the two types of methods to answer one and the same question. Quantitative methods then give the broad, generalizable, replicable overview, with qualitative methods shedding light on why the data have been observed as they are, what the processes were that generated the observed patterns. The advantage of such a 'mixed methods' approach is not only that we can answer questions about quantities as well as about how those quantities came about, but that each type of method in a sense 'buffers' the weakness of the other type. Such a combination of qualitative and quantitative methods in one study is encountered as 'triangulation', as 'mixed methods', or even as a multi-method study. Various types of mixed methods designs exist. An excellent and concise introduction to such types of research, which compares their strengths and weaknesses and unique suitabilities for specific questions, is Dixon, Singleton, & Straits (2016), chapter 4. Apart from combining qualitative and quantitative approaches in research, other examples of mixed methods exist, combining various quantitative data sources or different methodological approaches to answering the same questions. The terms 'mixed methods' and 'triangulation' are in practice often used interchangeably, although triangulation was originally coined to indicate the combination of qualitative and quantitative lenses on one and the same phenomenon or, more precisely, research question.

Zeng and Eleveld (2022) investigated, focusing on food riders in Amsterdam, the assumptions underlying the proposal for an EU Directive on improving working conditions in platform work (`ec.europa.eu/eures/public/eu-proposes-directive-protect-rights-platform-workers-2022-03-17_en`). Underlying the proposed directive is the assumption that riders wish to be employees rather than freelance, in order to be protected by labour law. For this study, the authors analysed the EU proposal, EU and Dutch labour law, and reviewed relevant literature. In addition, they conducted semi-structured interviews with 20 riders in Amsterdam. The researchers stratified their sample (see section 3.3.1: 'Stratified sample') in the sense that they conducted 10 interviews with riders who were employees and 10 with riders who were self-employed. They also observed popular rider areas, and conducted interviews with professionals such as experts, NGOs and the riders' union.

The authors found that most riders actually did not have much specific knowledge about labour law. Especially those riders who were employed already preferred employee status because it gave them stability and financial security. However, most riders preferred the flexibility and anonymity of being a self-employed rider. Particular

reasons for self-employment that had not been identified yet in the EU proposal were, amongst others, the possibility (especially for women) to refuse rides to neighbourhoods deemed unsafe, and the possibility to also cater for undocumented migrants.

2.8 Macro-, meso- and micro-level studies

The macro-meso-micro categorization is often encountered in sociological research where processes and explanations are regularly divided along these lines. The categorization reflects the aggregation level at which studies attempt to draw conclusions.

The *micro*-level is the lowest or smallest possible analysis unit. Often this is the level of the individual, and micro-level studies generally focus on respondents: people (whether they are defendants, lawyers, litigants, citizens) or cases. These are interviewed or observed, or data are collected about them from other sources. Explanations for the phenomena that are observed are then also situated at this level, for example an explanation where it is postulated that greed (an individual trait) causes people to embezzle.

The *macro*-level is found at the other end of the aggregation spectrum. As this is therefore the highest level of aggregation, explanations focus on high levels of aggregation, such as the level of the nation or government. An example of such an explanation is one where it is said that strict tenancy laws lead to a shortage of rental units.

In between the micro- and macro-level is situated the *meso*-level. Here, there is some aggregation over individuals. Examples of this level are neighbourhoods, companies, or trade unions. Explanations for the phenomena that are observed at the meso-level are then also situated at this level, for example the explanation where it is postulated that an open and non-hierarchical corporate culture prevents norm-transgressive behaviour.

2.9 Primary and secondary data

Many researchers collect their own data. They go out into the field, and ring doorbells with bailiffs, interview lawyers, survey citizens, code court cases, and the like. Such data, collected by researchers themselves, according to their own codes and formats and quality standards, are referred to as ‘primary’ data.

However, when doing empirical legal research, it is very often possible to use data that have already been collected by others. Governmental statistics agencies collect data on the number and nature of cases dealt with in court, the prison service surveys prisoners, law firms have administrations, the records of asylum seekers’ applications are held in governmental offices.

Such data, already collected by others, are often attractive to use. Firstly because of cost: if we can use governmental data or data collected by other bodies, we generally spend (if we have to pay at all) but a fraction of what it would cost to collect all that data ourselves. Second, such data are attractive because we then do not need to design and carry out sampling procedures: we may have access to all cases that exist.

In many countries, a wealth of data is available from registries held by national statistics agencies. For instance, for England and Wales the Individual Insolvency Register holds records on all bankruptcy and insolvency cases for a period of three months (`gov.uk/search-bankruptcy-insolvency-register`); Ireland and Scotland have separate registers). The extent of data and coverage over historical periods can be amazing. Statistics of bankruptcy and all civil cases have been collected annually for England and Wales since 1856, and are published and accessible online at Parliamentary Papers: for instance for all cases decided by the House of Lords, cases from late 1996 up to 2009 are available online from `publications.parliament.uk/`; from late 2009 onwards, all decided cases are accessible from the Supreme Court website (`supremecourt.uk/decided-cases/index.html`). Court data in England and Wales are well preserved, extensive and complete, and go back a very long time. Files of bankruptcy cases are held at the National Archives, with notices of bankruptcies from 1665 (!) to 1986; after 1945 the reports are generally held to become more summary. As these case files contain the original data, they can be considered primary data.

For the Netherlands, a selection of court rulings is published since 2016 at `rechtspraak.nl`; as only the rulings are published, these data on rulings can be considered primary data as they are simply the original rulings. However, information on lots of aspects of the case that is present in the original case file is mostly summarized, such as information on victims, or the pleas made by lawyers, or the different arguments used by plaintiffs. Information on the latter as it appears in the rulings should therefore be regarded as secondary data: registrars have summarized the data as far as is relevant for the ruling.

The sample of Dutch court cases at `rechtspraak.nl` unfortunately does not constitute a population (yet) in the statistical sense (see chapter 3). For instance, for criminal law, verdicts in all (completed and attempted) homicide cases are published, but an unclear selection of sexual violence cases is published. All cases decided upon by the Supreme Court are published, as are all cases decided upon by a number of higher appeal boards that are particularly relevant for business and commerce (inter alia the Administrative Law Division of the Council of State, the Administrative Court for Trade and Industry, the Business Chamber of the Amsterdam Court of Appeal, the Intellectual Property Division of the Civil Sector of the District Court of The Hague, and the Agricultural Tenancies Appeal Chamber of the Arnhem Court), that is to say, in as far as the case has not been declared unfounded or non-admissible or has been dealt with through a standard wording model. The Netherlands Council for the Judiciary has announced that it aims to publish *all* rulings within a few years, after which an entire population of rulings will be accessible online.

In all European countries, all decisions are given a unique European identifier (European Case Law Identifier or ECLI). ECLI is a uniform code that has the same recognizable format for all EU member states, composed of five mandatory elements, namely a country code, the code of the court that rendered the judgment, the year the judgment was rendered, and a number that is a unique identifier for the specific country. The Dutch decisions are searchable by this ECLI code, but also through other pointers, such as the specific court, the substantive area of law, or through more open terms such as 'tort', which produces numerous hits. The massive amount of hits is then refinable

(for instance by adding 'sports' as a more specific second term), which will reduce the set to be much smaller and more workable amount.

In many countries, the prosecution service holds statistics and court files. Until the end of the last century, most documents were paper only. Since the turn of the century, more and more of these case files are held only in digital format. It should be noted that not all data that are of interest to the ELS researcher are held by governmental bodies: for instance insurers hold interesting records (although they might not always be willing to share them), such as on all cases filed claiming legal expenses insurance. The latter would if available (obviously) not entail a complete list of all claimants or even conflicts, but again only a particular selection, namely those who submit a claim with their insurer.

Secondary data can be used only for some research questions. For other questions, we do need to go and collect data ourselves, because the government or other bodies simply do not collect the data we need, or do not collect them in the format or following the quality standards we require. It is therefore important to keep in mind that – while their easy availability may be seductive – there are drawbacks to their use.

2.10 Research ethics: rules and conventions

Whenever we conduct research in which we study humans, we have to observe ethical rules. Every discipline has its own – at times international – ethical rules, which in general share a common core. A first rule is that research subjects or respondents have the right to be treated decently and respectfully. Appointments should be met, on time, and researchers should dress and behave according to the standards of the respondents. The same goes when we approach organizations whose data we might want to use. Attentive e-mails and swift responses generally get you further than informal or cursory responses. And, apart from this being common decency, experience teaches that treating respondents with respect, friendliness and flexibility is also much better for the research: respondents who feel they are not being treated correctly will maybe sooner refuse to participate, or may give only cursory or half-true answers, just to be rid of the researchers or to quickly cash the remuneration.

Second, and more formally, respondents have the right before they agree to participate to be informed what the research is about, who finances it, what will happen with the information they give – in short, they are entitled to be fully informed about all aspects of the research that may be relevant for them in deciding whether they want to participate or not. This is referred to as *informed consent*. We return to this topic in section 4.2. Obviously, researchers should also heed the safety of respondents. Especially in charged environments, such when studying conflict settings, the participation in a study may by itself endanger respondents.

In the following paragraphs, we discuss the legal framework for data protection and research on human research subjects. We describe how internal and independent review boards commonly review research proposals before any research goes ahead, and we delve a little into the implications for data protection and ethical rules when research is conducted online.

2.10.1 Legal framework: the GDPR

The European Union General Data Protection Regulation (`eugdpr.org`) constitutes the legal framework that governs the manner in which the privacy of citizens and therefore also respondents (in general: non-deceased human study objects) should be safeguarded. Because of its relative novelty (it came into force in 2018), it is not yet clear what the exact scope is of what the GDPR prescribes and forbids when it comes to data protection. All research institutions are required to have a data protection officer (DPO) in place, an independent data protection expert responsible for monitoring compliance, informing and advising on data protection obligations, and who is able to help with practical and technical questions. It does appear, from recent experience across the social sciences board, as if procedures have become much more complicated and lengthy, and as if those who need to give permission for the use of data have become much more strict and reticent in granting access to data.

While much is still unclear about the exact implications of the GDPR for scientific research, it is clear that every researcher is obliged to carry out a so-called Data Privacy Impact Assessment (DPIA) before an investigation can go ahead. A DPIA describes the handling of (special) personal data (`gdpr-info.eu/issues/privacy-impact-assessment/`) collected during the research. A DPIA should specify the risk of the investigation for the rights and freedoms of the persons being examined when processing their personal data, and the measures taken for concrete protection.

In doing so, it is helpful to distinguish three types of data: 'red' data (containing personal details or data that can be directly traced back to persons, such as their name, address, date and place of birth), 'orange' data (that do not contain directly personally identifiable data, but data that in combination with other data can reveal the identity of the persons being investigated; this data is also referred to as 'pseudonymized'), and 'green' data (that contain absolutely no personal or personally identifiable data).

Red data is the most sensitive from a privacy point of view, green data is completely insensitive, and orange data is in between. The regimes that apply to the storage and transport of this data are correspondingly strict. Green data files may be transported freely (physically on paper or on a USB stick, or via e-mail), because they cannot violate the privacy of persons. Red files should preferably only be stored on a data carrier that cannot be accessed or is extremely difficult to access by anyone other than the researcher. In more and more institutions, this type of data is stored on a standalone workstation that is not linked to the internet ('air-gapped') so that it is technically 'unhackable' from outside. If red data need to be transported, this is done on a special USB stick that is encrypted. Orange data can be stored on an ordinary workstation such as a laptop or PC (which is therefore connected to the internet) but may not be transported freely.

An often-used practical solution for transporting data collected from individuals in the field is to divide the data into two sets: one set that contains the personal data with a unique code per respondent, and another set that contains the unique code and the research data. The unique code is the key with which the sets can be linked and research data linked to individuals. Transported separately, the two datasets are fairly harmless if they were lost: the first set contains only names and meaningless numbers,

and the second meaningless numbers and research data that cannot however be linked to individual respondents.

In general, we recommend that one be too careful rather than easy-going. It is also advised to be extra-careful when conducting research abroad or in collaboration with researchers from other countries. Special agreements must sometimes be laid down for cooperation with certain countries, and in some countries special privacy legislation also applies.

The French Centre National de la Recherche Scientifique (CNRS) has drafted an overview of the GDPR and its implications for research in the humanities and social sciences (downloadable from `inshs.cnrs.fr/sites/institut_inshs/files/pdf/Guide_rgpd_2021_en_0.pdf`; CNRS, 2021) that explains key concepts in the GDPR and describes what they mean in practice. Although geared towards the French situation and the CNRS, and with particular focus on health data, it gives useful examples of consent forms and information notices, and ends with a useful checklist of key issues to achieve compliance with the GDPR on the protection of personal data.

2.10.2 Code of conduct: the ALLEA Code

Since 2017, a European Code of Conduct for Research Integrity has been adopted, a revised version of an earlier code. The Code is quite readable and lists principles, good research practice, as well as violations of research integrity. The principles are briefly summarized as: *reliability* in ensuring the quality of research, reflected in the design, the methodology, the analysis and the use of resources (very much relevant for the topics discussed in this book), *honesty* in developing, undertaking, reviewing, reporting and communicating research in a transparent, fair, full and unbiased way, *respect* for colleagues, research participants, society, ecosystems, cultural heritage and the environment, and *accountability* for the research from idea to publication, for its management and organization, for training, supervision and mentoring, and for its wider impacts.

The Code categorizes fabrication, falsification and plagiarism as research misconduct, and lists a number of unacceptable practices (such as citing selectively to enhance one's own findings or to please editors, reviewers or colleagues, withholding research results and expanding unnecessarily the bibliography of a study), and subsequently lists a number of principles that national institutions and research institutes should adhere to in dealing with violations and allegations of misconduct.

The Code addresses emerging challenges emanating from technological developments, open science, citizen science and social media, among other areas. The European Commission has recognized the Code as the reference document for research integrity for all EU-funded research projects and as a model for organizations and researchers across Europe. The Code was published originally in English on 24 March 2017 and was translated to all official EU languages by the European Commission's Translational Services and with the support of ALLEA Member Academies. See `allea.org/code-of-conduct/`.

2.10.3 Internal review boards

Many research institutes and universities call in the aid of independent or internal review or research and ethical boards for all research in which human subjects are involved. Such committees are hugely important, and are more and more a standard precondition from management to give studies the green light. Not only do they give independent 'clearance' for studies to go ahead, they also constitute more and more often a general requirement for research funding to be released. In addition, they can also be extremely helpful in thinking about the mitigation of risks and the protection of respondents and staff.

Internal Review Boards (IRBs) are thus necessary and invaluable steps in the design and preparation for studies. IRBs will screen study designs and protocols to see whether informed consent is guaranteed, whether respondents' safety is ensured, whether staff safety is secured, and whether any breaches of principles are proportional.

It should be noted that while research ethics generally refer to respondents, they are equally important to consider for researchers. When research takes place in charged or volatile settings, the safety of the researchers is an issue that needs consideration too. Insurance constraints may decide whether research can go ahead, for instance in interviewing forensic staff in conflict areas. But even if insurance covers the researchers, it is important to consider whether one needs or wants to expose interviewers to the risk of getting hurt. IRBs therefore generally also take into account the safety of research staff when assessing research proposals.

Van den Berg, Blommaert, Bijleveld, & Ruiter (2020) investigated the effect of a criminal record on job market opportunities by sending motivation letters with CVs in response to existing vacancies, with one half of the applicants randomly 'confessing' that they had a criminal record for a certain offence, and the other half not. In this study, the employers did not know that the applications had been fabricated and sent by researchers. They participated unknowingly, and without informed consent, in a field experiment. Nevertheless, the exception to informed consent could be defended here, because informed consent in advance would have made it impossible to answer the research question: the employers would then no longer behave 'naturally' because they knew they were being studied and what the focus of the researchers was.

The researchers found, to their surprise, that a criminal record had but a marginal effect on applicants' chances of getting back a favourable reply. The effect of the indicated migrant background of the applicants (the researchers had also randomly varied the names of the applicants, one name indicating a Dutch applicant, and the other name indicating a migrant background) was much stronger, however, with a Dutch applicant who had confessed to having previously committed a sex offence significantly more likely to receive a positive reply than an applicant with a migrant background with a clean rap sheet.

Flood (2005) conducted an ethnographic study in a law firm in Chicago but did not go 'undercover'. He did introduce himself to all staff as a researcher, which takes away any objections regarding informed consent: all firm employees knew who he was, what he was studying and were therefore able to choose – to a certain degree – to what extent they wanted to interact with him. Revealing that you are a researcher and what and for whom you are studying, generally takes away any objections an IRB

might have. Obviously, the downside for the Flood study might have been that staff members may have been less eager to talk with him or show him their true feelings or thoughts, let him in on all that was happening at the firm; the research was therefore less *unobtrusive* (a term we will return to in section 4.5), so that validity may have been at stake.

Chapter questions

1. How do the empirical cycles of quantitative and qualitative research differ? (section 2.2)
2. What is the difference between a conceptual and an operational definition? (section 2.4)
3. Look up the definition of legitimacy in political science and the definition in legal scholarship and describe their respective conceptual scopes. What are the implications of any different conceptual definitions for the respective operationalizations? (section 2.4)
4. How are reliability and validity connected? (section 2.6)
5. How is reliability generally assessed? (section 2.6)
6. What types of validity are distinguished? (section 2.6)
7. Give examples of micro-, meso- and macro-data (section 2.9)
8. What is meant by informed consent? (section 2.10)
9. Discuss whether it is unethical when participants in a study have not given full informed consent (section 2.10)
10. Outline the main areas of overlap and difference between the GDPR and the ALLEA Code (section 2.10.1 and section 2.10.2)