Methods for Empirical Legal Research

Catrien Bijleveld

Abstract

This chapter is part of a book that focuses on new EU legislation in areas that link to digitalisation, such as the DSA, the EMFA, and more generally the GDPR. Much of this legislation is relatively new. And most of it is fairly extensive and complex, with, for instance, the DSA (English version) comprising more than one hundred pages. It is therefore more than laudable that the editors of this volume have chosen to bring together scholars to facilitate understanding and research into this legislation.

This Chapter will briefly describe some core principles for carrying out empirical legal research. The introduction to commonly employed empirical research methods will be basic and conceptual. In this chapter, I borrow from Bijleveld (2023), which provides a more extensive, yet easily accessible and conceptual, introduction to Empirical Legal Studies (ELS).

1. What are empirical legal studies

Empirical legal studies, or empirical legal research, is a label given to studies that focus on the law by gathering empirical facts. ELS is, in a sense, a subfield at the fringes or the intersection of law and social sciences. It is also encountered as *empirical legal research* or *legal realism*/*new legal realism*. Similar – but not exactly identical – areas of study are denoted as *law in action* or *legal sociology*. Some (sub)disciplines share properties with empirical legal research: criminology does, and so do legal sociology, law and economics, and legal anthropology.

Questions addressed in ELS all inquire into empirical facts, and can be categorized into three pillars or a "trias ELSica" (Bijleveld, 2023). They focus on the law's assumptions (such as that harm can be repaired by monetary compensation), its operations (such as the time it takes to reach a decision in court cases), or effects (such as whether the DSA is successful in protecting consumers and their fundamental rights). The pillars are intrinsically related. If the assumptions on which laws are built are incorrect,

or if tradition or lack of intrinsic support for new rules stand in the way, then it is very unlikely that the laws would have their desired effect. If the assumptions are correct, but the law is not applied as planned (for instance, cases take extraordinarily longer to process, and judges find the new rules unworkable), it is also unlikely that the foreseen effect would materialize. In that sense, the three pillars form a trias.

ELS is clearly not doctrinal. In doctrinal research, case law, or the extent to which laws and regulations are in line with treaties or supranational law, are studied (see, e.g., Hutchinson, 2013; 2015; Van Boom, Desmet and Mascini, 2018; Van Gestel and Micklitz, 2011). For instance, in jurisprudence analysis, we may be interested in how cases have been dealt with, what arguments have been used to find accused parties liable, and what the threshold is that the Supreme Court employs for finding that there was criminal intent. Scholars who analyse such case law do so in a fairly targeted way. They pick the exemplary case law to prove a point or illustrate a new turn in evidentiary practice. However, it is very conceivable that one legal scholar would arrive at a different conclusion when investigating the same doctrinal issue simply because they regard different cases as pertinent or adopt a different philosophical stance. It then also depends very much on the scholar's authority to what extent the conclusion is regarded as valid.

Contrary to doctrinal research, data collection in ELS is done systematically, according to a well-described and accepted set of rules. In ELS, we gather empirical facts about the law and investigate what occurs in the measurable world around us, what happens within legal practice, and what the effects of laws are. We want the person collecting the relevant facts to serve solely as the vessel through which the data are presented, forming the basis for the conclusion. Formulated conversely, we would not want our understanding of the empirical world around us to depend on what particular scholar carried out the research. The real world is out there, and we would want each scholar who employs the same systematic empirical approach to arrive at approximately the same conclusion about that reality.

However, while ELS is empirical and appears disjunct from doctrinal research as different questions are asked in ELS and different methods are used, the core interest of all ELS *is* the law. What distinguishes ELS from other empirical disciplines such as legal sociology, law and economics, or legal anthropology is that an ELS scholar will always want to translate their findings back to the law. What do the empirical findings mean for how laws have been drafted? What do they mean for legal practice? The core

interest of the ELS scholar is the law and not the testing of an economic or sociological theory.

ELS is, therefore, not disjunct from legal, doctrinal research. Davies (2020) argues that the doctrinal and empirical study of law should, in some way, enrich each other. Van Boom, Desmet and Mascini (2018, pp. 5–6) write that the empirical study of law enriches doctrinal legal research beyond empirical fact-checking because it allows a deeper understanding of not only the plain facts but also the underlying mechanisms of legal interaction, including insight into both explicit reasoning and unconscious processes in legally relevant decision-making. ELS is, therefore, part and parcel of the legal discipline, sometimes indicated by a hyphen as in: *empirical-legal studies*.

What is important at this stage is to note that, from the various definitions, three defining characteristics of empirical legal research emerge, namely that (1) an empirical legal study poses questions about the law, (2) it systematically collects empirical data to answer those questions, and that (3) the answers to the questions are legally relevant. What precisely the latter is remains fairly vague. In general, we mean by this that, in some way, we would want to be able to translate back the research findings to the law and legal practice. For instance, a study might find that an applicability test for social benefits has been formulated so vaguely that wide discrepancies exist between officials in interpreting these norms, thereby threatening equality before the law. The study could then point out that clearer norms or criteria need to be formulated.

Given that ELS evolves around the analysis of the empirical world, empirical methods are used. Empirical methods are used in many empirical disciplines; in that sense, these methods are not particular or new. We encounter both quantitative and qualitative methods. Across the quantitative board, we find univariate methods, such as means, medians and percentages. We find correlational methods, such as simple correlation measures and cross-tabulations, that give a feel for association through chi-square measures and odds ratios. Multivariate methods are mostly used first when we want to predict an outcome from a set of characteristics. Regression analysis can be used to predict sentence length from gravity of the crime, mitigating and aggravating circumstances (see, for instance, Hola et al, 2015). Analysis of variance or ANOVA is often used when we analyse data from vignette studies, where the variables have a specific format. Other, less run-of-the-mill multivariate techniques may be used, such as factor analysis

or multiple correspondence analysis, to identify risk profiles of persons placed under guardianship measures, looking for particular combinations of mental and physical health problems, financial problems and issues in their support network (Nieuwboer et al, 2025). We sometimes encounter methods from other disciplines as they are particularly suited to the type of data we collect for the phenomena we are studying. A technique borrowed from epidemiology, for instance, is generally used if we study disposition times: we then need so-called time-to-event or survival methods (see, Bijleveld, 2023, chapter 6), for which analysis techniques can be univariate, bivariate or multivariate depending on the complexity of the models we are using. Quantitative methods, specifically econometric methods, are almost universally used in Law and Economics, which clearly overlaps with ELS.[1]

Qualitative methods are also widely used, amongst which the most prominent probably is content analysis. It is used to analyse textified material, such as court files, applications, and interview transcripts. Qualitative methods are very flexible and can also be used to analyse behaviour that has been systematically observed (such as courtroom interactions) or captured in video material (such as CCTV-recorded interactions between officials and citizens). The analysis of such materials can be done deductively, that is, departing from a given theoretical framework in which the researcher investigates to what extent certain characteristics are present, or inductively, in which case the researcher approaches the material and seeks for patterns, repetitions and so-called *themes* in the material (see, Bijleveld, 2023, Chapter 7, and Tracy (2013) for textbook introductions, and, specifically on qualitative methods in empirical legal studies, see, Webley (2010)).

Sometimes, particular data collection methods are used because of the nature of the phenomena being studied. For instance, if we are interested in estimating the prevalence of fraud, we must account for the fact that respondents may not be eager to divulge behaviour that they are ashamed of and we may need to employ specific methods for sensitive topics, such as randomized response (see, John et al, 2018, for a non-technical overview, and for some examples, see, Bijleveld, 2023, Chapter 10).

1 The US based Journal of Empirical Legal Studies has numerous examples of the application of such methods. Additionally, Chapter 38 in Cane and Kritzer (2010) gives an overview of quantitative methods in ELS.

2. Doing empirical legal research

If we want to study the assumptions made within the law, its operations, or effects, we mostly use so-called constructs. Constructs are variables that are considered relevant for the research (such as *trust*), but that are not directly observable (which a variable such as *sentence length* would be). Given that a construct is not easily observable, a definition must be given, and it needs to be *operationalised*, it needs to be laid down how exactly we could measure it.

An example of such a construct is *procedural justice* (Tyler, 1990), which 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 accordance with fairness principles, they are more likely to comply with the outcome, even when the outcome 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 considered 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 can 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).

Operationalizing is sometimes not straightforward and is relatively easily exemplified with the psychological construct *intelligence*. While the word intelligence is common usage in many languages, the 1981 version of the Wechsler Adult Intelligence Scale, for instance, presented verbal and performance-scales, measured with several subtests, five for verbal and six for performance abilities. Other instruments (such as the Raven test, which is nonverbal) use a conceptual definition that differs or operationalizes intelligence (slightly) differently. If one uses a different conceptual definition of intelligence or a different operational definition, intelligence measurements will differ across definitions. The same goes for constructs used more often in ELS, such as trust or justice. An application of operationalization in the

study of medical malpractice can be found in Van Velthoven (2016), and a nice illustration of how different operationalizations unpack in practice in Haucke, Hoekstra and Van Ravenzwaaij (2021).

Constructs should be operationalized to ensure they provide both a valid and reliable measurement of the property investigated. A *valid measurement* is a measurement that truly, validly represents the property of interest. For example, an intelligence test that measures only arithmetic skills does not represent the entire spectrum of what we suppose intelligence contains. It will produce an invalid measure of our construct intelligence. A test that is very verbose will not be able to measure the intelligence of recent migrants who have not yet mastered the local language. We would also like our test to predict (to a certain extent) school success, as we expect performance to correlate with intelligence. A measure that does all that, we label as *valid*. Comprised in validity is the idea of *reliability*, as the measures should be precise. Again, a counter-example of what we mean by reliability is the following: using an elastic measuring tape would, for instance, make for an unreliable measure of people's height. One time, a person's height would be measured as 170 cm, next as 172, then 167, then 173, etc. If the measurement was done 100 times, the result would likely be, on average, right. However, the measure is considered unreliable because of the variability in the measurements.

A reliability check is often done by having two observers code the same feature independently. If the results from these two raters concur, *interrater reliability* is present. Reliability can be expressed as percentage agreement as well, and other measures exist. Assessment of validity is more complex, although, in general, face validity is often employed (essentially, whether the measures look credible and in accordance with the definition). See Bijleveld (2023, Chapter 2) for a succinct overview and Drost (2011) for a more extensive treatise focused on psychometric research.

Validity and reliability are important in themselves, and also because scientific research needs to be *replicable*. For important conclusions on the operations of the legal system or the effects of the law to be solid, and not a one-off result, we want them to be corroborated by several independent researchers. As said above, different researchers using the same instruments should arrive at roughly the same conclusion about the world. Valid and reliable findings build confidence in the relevance of the findings, and provide a basis for evidence-based policy. Validity (and inherent to it: reliability) is a necessary condition for research to be replicable.

It is not a sufficient condition, however. Mainly for psychology and health research, a *replication crisis* has been identified. Studies have been repeated with the same definitions, measurements, and procedures, but rendering different results. Such different results are, of course, highly problematic. Replicability should not be confused with reproducibility, which is generally understood as different researchers analysing the same data and arriving at the same result. Both reproducibility and replicability are important desiderata, and increasing focus is put on encouraging (or even requiring as a condition for funding) that researchers make their datasets available for re-analysis by others.

3. Empirical legal research: qualitative and quantitative methods

An often-used categorisation of research that we already briefly touched upon is the division into qualitative and quantitative studies. Formulated simplistically, the two can be characterised as follows: while quantitative studies aim to measure the volume or *quantity* of some variable of interest, qualitative studies are geared towards discovering the *quality*, *nature*, *why* or *how* of phenomena.

In a study that uses quantitative methods, the goal is to understand *how often* something occurred, such as: "How often are cases of domestic violence acquitted?" or "What percentage of citizens have trust in the criminal justice system?" or "How many citizens with a certain type of legal problem take their case to court?" Quantitative studies typically follow a fairly strict format (the empirical cycle) in which hypotheses are formulated and where statistical testing is generally employed. Also, samples are generally large in quantitative research, and standardised instruments (such as coding lists or web surveys) are often used. An explicit aim is to generalize findings from the studied sample to a larger population. On the other hand, qualitative methods are used to understand *why* things happen or *how* and to explore new phenomena. Examples of questions we would pose then are: "What are the reasons for taking or not taking a business conflict to court?" or "Under what circumstances are domestic violence filings settled through mediation?" or "What deliberations do judges make in divorce procedures when one parent has accused the other of sexual abuse?" Qualitative studies are generally much less strictly formatted beforehand than quantitative studies. Hypothesis testing is rare, and statistics is therefore used much less often. Qualitative designs differ from quantitative methods: smaller,

not necessarily representative samples are generally used. Open interviews, focus groups and observation are common, and analytic methods are less prescribed and more exploratory, often spread over several iterations.

Quantitative research generally produces a broad, generalizable, quantitative summary of a phenomenon. Qualitative research gives a rich understanding of a particular problem within a particular context. As the two are different methods for answering seemingly different kinds of questions, neither is superior to the other. The adverb *seemingly* is not used without purpose, however, as many questions can be addressed using either quantitative or qualitative methods. The approach may then be different, depending on what type of methods and answers are chosen.

In qualitative research, the aim is much less to produce generalisable quantitative statements but to unravel a number of mechanisms, to *understand* what happened, or to understand the meaning that the research subjects give to the phenomena being studied. As qualitative scholars work from the assumption that all human enterprise is contextual, they tend to study phenomena, and understand phenomena, within a given, particular context. Therefore, qualitative research is inherently less generalisable.

Quantitative studies are sometimes irreverently qualified as shallow. In a quantitative study, only a few factors or variables are investigated. Contextual effects are generally not included but seen as a nuisance: quantitative researchers attempt to isolate the variables they are interested in and control for any contextual noise that might distort the picture. Examples of such studies are experimental studies into the effectiveness of medicines. A group of patients is selected, and the medicine to be tested and a placebo are administered randomly among the group. Any differences between the group that received the medicine and the group that received the placebo are then attributable to the medicine and the medicine only. In such a design, the medicine is *isolated*, and the impact of any contextual effects (such as the expectations patients had, any other medical conditions patients have, their gender, or personality characteristics) is evened out by randomisation.

In summary (and admittedly leaving out nuances), a qualitative study picks a small part of the population of interest, but it delves deep, goes to the bottom of things and generates a rich and contextual understanding. However, whether the same result would have been found elsewhere cannot be guaranteed, as the findings apply only within that particular context. A quantitative study looks at a few aspects of the problem at hand but does so broadly and tries to find the impact of factors regardless of any particular context. That makes the findings of a quantitative study more easily gener-

alisable across contexts. As it largely disregards context, it investigates only a limited number of aspects of the problem at hand.

Why is it important to touch upon this distinction? Because the two traditions or paradigms use partially different methods. Qualitative studies rely more on open interviews, analysis of texts, observations, and immersing oneself in the context to be studied. Samples are generally smaller. Studies can be planned only to a certain extent, as it is uncertain beforehand what will be encountered. The analysis is generally lengthier and iterative. Quantitative studies, on the other hand, rely more heavily on pre-designed measurement instruments, such as scoring protocols or web surveys. Extensive piloting is necessary. Samples are generally larger, and testing, model building and statistics are common.

Many ELS students prefer qualitative methods to quantitative, assuming that qualitative research – without maths and formulas – is easier. The latter is, however, generally not the case. Qualitative research requires strong theoretical skills, hard and good analysis, and perseverance, constituting more often than not a substantive investment (and may entail much more – tedious – work than quantitative research). Whether the outcomes are useful is also often more uncertain beforehand. A solid qualitative study is a feat that requires extensive training and is much harder to learn through textbook recipes which can be used for teaching quantitative skills.

However, what many scholars recommend, and this author is one of them, is to combine the two types of methods whenever possible. As each type of method has its drawbacks, using both types can help to buffer the weaknesses of one through the other. If two different methods are used to answer the same question, this is called *triangulation*. By using multiple methods, we do not rely on one technique only, allowing more confidence in the research findings, their credibility, and their validity. Studies that use multiple methods are also referred to as *mixed methods* studies. Both terms (triangulation and mixed methods) are also used when researchers use different datasets; here, too, the idea is that by not relying on one source of data only, we can be more confident of the findings.

4. Sampling, representativeness and testing

As in all social science research, empirical legal research often involves working with samples due to limited time and resources. A population can be a population in the literal sense, such as all European Union inhabitants,

or all defendants at the International Criminal Court. A population can also consist of non-humans, such as all cases filed at a certain court or all verdicts in homicide cases. The population is the universe of units the researchers are interested in and want to draw conclusions on.

If only a part of that universe is studied, our knowledge of it is incomplete. As not all population members were studied, no assurance can be given that the sample results also pertain to the entire population. While sampling only a part of the population saves a lot of expenses, the flip side of the coin is that in doing so we have introduced *uncertainty*. We are unsure of what is called *external validity*, that is, whether our conclusions about the sample also hold true for the larger population.

However, scrutinizing each and every population member is actually not necessary. By following certain rules and with reasonable precision, conclusions about the entire population can be drawn, even if only a part of it, a *sample*, is investigated. Often, a small part will already do, like a 1% sample, or even less, depending on various factors. Statistics is the science of dealing with the uncertainty that sampling introduces. It provides the rules and procedures and the means to infer levels of uncertainty – or, conversely, confidence – about the conclusions drawn from the sample regarding the population.

4.1 Sample representativeness

A sample's properties resemble the population's properties. In statistics-speak, we want a *representative sample*. The easiest way to ensure that a sample's properties reflect those of the population is to draw that sample by chance or *at random*. In that case, every population member has an equal chance to be part of the sample, which is now called a *probability sample*. For that, a list of all population members is created (the so-called *sampling frame*), population members are numbered, and the desired number of sample members is chosen using some random number-generating tool. When studying case law, for example, a list of all court cases could be compiled, and a *random sample* using such a tool could be drawn. Or, if 20,000 cases are accessible and sufficient time and funds to analyse 500 cases, a random number between 1 and 20,000 is picked, and we sample every 40th case. This is called a *systematic sample*. Another option is to employ a so-called *cluster sample*: in ELS, we often find cases dealt with at different district courts within one country. One could now first draw

a random sample of courts and then, within each court again, a random sample, saving the trouble of having to collect data at each and every court. Such cluster samples are pragmatic but come at a methodological cost (the "design effect", see Bijleveld, 2023, chapter 3).

In practice, however, non-probability samples are often drawn due to a lack of sampling frame, lack of access or resources to go through all the motions of random sampling. While one should always strive for random sampling, non-probability samples may, in fact, be quite useful. They may even be representative, but representativeness is not *guaranteed*. In qualitative research, non-probability samples are frequently used. For instance, interviewing a sample of professionals who were chosen because they have specific expertise in the observation of interactions between parties involved in conflicts dealt with in a court.

In ELS, it is often technically possible to study entire populations. It may be that case law is available online, or all defendants or all litigants can be studied because case files have been digitised and electronic databases are (under some conditions mostly) available for research. The increasing digitisation of case law is a very attractive outlook for ELS. For the near future, practical constraints will make many scholars still resort to sampling, as it may be too time-consuming to study massive amounts of data, even if they have been digitised. However, as more software becomes available for automated text analysis, it is likely that enormous amounts of textified material and in fact entire populations of case law can be analysed (Dyevre, 2021).

4.2 Sample nonresponse

In most practical situations, sample nonresponse occurs, meaning the selected members cannot be assessed or sampled. This is firstly so during citizen surveys. Depending on the topic of the study, the infrastructural possibilities, the persuasive skills of interviewers and the like, survey response rates generally hover between 20% and 40%; higher response rates are rare. Therefore, to aim for a sample of 100 respondents might lead to only 40 completed interviews, a so-called *retention rate* of 40%, and an *attrition rate* of 60%. One might be tempted to think that this is not a real problem, as a larger initial sample of, say, 250 could be drawn, and then the target of 100 interviewed respondents could be reached. Unfortunately, this does not solve the problem that nonresponse generates. The problem

is namely not simply that the survey has fewer respondents. The problem is that nonresponse is generally not accidental, not random, as it is not a coincidence that certain respondents do not end up in the realised interviewed sample. Often, the vulnerable and the elderly who are too ill to be interviewed, the mistrustful, those who are afraid to talk to strangers or the busy bees with 80-hour work weeks refuse to talk to researchers.

Even if the research starts with a randomly drawn list of sample members, the non-random attrition process will lead to a non-random selection of the original random sample. Formulated more loosely: nonresponse messes up the representativeness of a sample. One might be tempted to think that this is a particularly problematic phenomenon when doing surveys with people in person who can be ill and who may decline. Attrition, however, also plays a role when studying, for instance, court files or treatment dossiers. Court files of defendants who have their cases up for review are typically unavailable and not to be found in the archive. The treatment files of recidivists may have been requested for inspection by the investigating psychiatrist or psychologist. Dossiers of withdrawn claims are cleaned earlier than those of cases taken to court. Thus, also here, the particular, atypical files will be missed, and a non-representative part of the original sample will be left for inspection.

Nonresponse is essentially irreparable. One can inspect the resulting sample thoroughly with a so-called *nonresponse analysis* and hope it resembles the population on pertinent characteristics (if known), such as age, gender, type of claim, geographical origin, and the like. If there are no serious differences, that is, if the realized sample resembles the population on such background characteristics, then that is more comforting than if differences were found. However, this background variables check does not contain information on whether the non-responders differ from the responders on the key variables of interest central to the main research question.

Nonresponse rates vary per topic and per type of study object (paper or electronic sample members, such as case files, generally do not generate high nonresponse rates). But nonresponse rates can be so high that generalisation to the population becomes increasingly unrealistic. Especially when the topic is sensitive, response rates as low as 2% have been encountered. Response rates of 40% to 50% are generally perceived as acceptable, even though then one should always check to what extent the non-responders differ from the responders.

4.3 Testing

When quantitative research is conducted, statements such as "this result is significant" or "regular divorce procedures take significantly longer than procedures with mediation" are often made. What is meant by such statements? While a detailed explanation will not be provided here, a brief overview of the concept of statistical testing will be offered.

Take the following example. After drawing a random sample of court rulings in cases of robbery, it can be seen that female defendants are handed down lighter sentences than male defendants. That might be not only the case in that sample but also in the population of all court cases. While confidence about the observation in the sample is high, in fact we are certain of the sample result, certainty about this *generalisation* cannot be postulated, as the entire population could not be observed.

Drawing a sample is a chance phenomenon, so could not the finding be simply attributable to chance, a random result, or coincidence? Because the sample was drawn randomly, the population might be reflected, but even so, uncertainty does remain. In order to deal with this uncertainty, statistical tests are used. There are very many different kinds of tests. However, the basic rationale of these tests is always the same. And this rationale is not difficult, as it follows the kind of reasoning each of us applies in everyday life.

Basically, the reasoning behind statistical testing is as follows: It begins with an assumption about the phenomenon we are interested in drawing conclusions about. Suppose, as an example, that we aim to investigate whether a new divorce procedure that includes mediation makes for shorter conclusion times than the standard divorce procedure. The assumption at the start would be:

H_0: $T_{old} = T_{new}$

In words, the new procedure takes just as long as the old procedure. This assumption is also called the null hypothesis: there is a null effect (also: H_0). We also formulate an alternative assumption, the alternative hypothesis, that is:

H_1: $T_{old} \neq T_{new}$

In words, the times to the conclusion of the new procedure and the old procedure differ. This assumption is also called the alternative hypothesis (also: H_1).

Now, assuming H_0 were true, we calculate the chances of finding our sample outcome. Suppose that that likelihood is very small, in order words: it is really unlikely to encounter such sample findings if H_0 were true, we then no longer assume that H_0 is true and we conclude that H_1 must be true. So, we conclude our findings are incompatible with the conclusion times being equal and the two divorce procedures' conclusion times differ.

While the statistical process may appear very abstract, as said, it is exactly the reasoning used in daily life. For instance, tossing a die 10 times, and each time finding the result of the toss being a six, would lead to the conclusion that the die is not fair. Eating at a canteen several times and falling sick each time would lead to the conclusion that unhealthy food is served there. In both examples, one is not 100% certain that this is the case. It is possible for a die to be tossed 10 times and each time a six ending on top, or, coincidentally, dinner at that canteen may coincide with a flu wave each time. Without measuring the die with a nifty device to see whether it is balanced, or without looking for bacteria in the restaurant food in a petri dish, we are not 100% certain of our conclusion.

We simply find it *too coincidental*. We accept a small risk to draw a wrong conclusion, namely that we conclude that H_1 holds, while actually H_0 is the case. That risk is called the *significance level*. Given that H_0 is formulated as the situation where nothing out of the ordinary is going on (no effect, no difference), this small risk – the significance level – is the likelihood of wrongly concluding that something interesting is going on when actually there is no effect or no difference (a false alarm). Significance levels of 5% are often regarded as acceptable, although this essentially depends on the risk a researcher wants to take in drawing a wrong conclusion here.

A researcher may wrongly conclude that H_1 is true while H_0 is actually true. But the opposite can also occur. If one is very risk-averse and sets the significance level very low (for instance, at 1% or 0.01%), one will simply never reject H_0. If a six was tossed 100 subsequent times and only then the unfairness of the die is assumed, one is so strict that one will almost never be able to conclude that the die is not fair. The test then has low *statistical power,* or briefly, low *power*: it is unable to detect that something out of the ordinary is going on. The power of a test is defined as the chance to decide that H_1 is true if it is true.

A good example to illustrate why statistical power is also important is a fire alarm. A fire alarm is calibrated to sound the alarm above a certain threshold of particles in the air. So, in that a sense it is like a statistical test. It cannot see whether there's a fire. It derives conclusions from sampling the air. Above a certain threshold, it will conclude that there is a fire and start sounding; below, it will remain silent. A false alarm can be very annoying. If one were to fiddle with the threshold (reducing the likelihood of a false alarm) this will the alarm to start screeching less soon. One then however increasing the likelihood of missing out on a fire, something much more problematic than annoying. The latter is the analogue of low statistical power: setting the significance level so low that one does not detect what is going on.

A large sample provides – ceteris paribus – larger power. In general, the chances of drawing the wrong conclusion on the population of interest are reduced when using a larger sample. This is quite logical. If one draws a larger sample out of the population of interest, one has observed a larger chunk out of that population and is therefore surer about what is going on in that population. This can be shown mathematically, but it is also intuitively so.

Much more can be said about statistical sampling. There are numerous kinds of tests and different ways to construct the null and alternative hypotheses, but for sake of brevity in this Chapter, I refer to general statistical textbooks and the non-technical introduction given in Bijleveld (2023, Chapter 7). Importantly, all testing follows the same rationale outlined here, and that that rationale is one we also often use in daily life.

5. Causality

In empirical legal studies, many questions centre around the impact of laws. Are cases concluded more swiftly because procedures were changed? Are rents down because of the new law restricting the rent that rental agencies may charge through a tariff system? Do female defendants get lighter sentences because they are female? These are causal questions. In each example, one would want to know not whether cases are concluded more swiftly before and after a law change but whether they are concluded more swiftly *because of* the law change. In the second example, it is not sufficient to establish that rents went down, what the research question points to is whether that was due to the new tariff system. For the last example,

the research question cannot be answered by establishing whether women receive lighter sentences than male defendants but it must be established whether that is due to their gender.

Pursuing the last example, a simple comparison of sentence lengths for men and women will not answer the question of discrimination. Men and women might commit different crimes, and this difference, in fact, explains any difference in sentence length. Even if we would compare sentence length for men and women within one type of crime only, different so-called *confounders* could be at play, making it impossible to infer anything about the effect of gender on sentence length. For instance, female defendants might be more remorseful, or more male defendants have a criminal record already, which translates to a heftier sentence for them.

The gold standard for assessing causality in empirical research is through an experimental design, where one randomly chosen set of research objects receives some kind of intervention, and another randomly chosen set does not. This type of design is often found in pharmacological research, where questions about whether medicine reduces complaints or vaccination protects against disease are determined. However, simply administering the intervention of interest to one group (the experimental group) and not administering it to the other group (the control group) is not enough to assure that any difference between experimental and control groups is attributable to the intervention. To make the experiment successful, the persons in either group should not be aware of which group they have been placed in, which is usually achieved by administering an empty intervention to the control group (a placebo). The COVID-19 vaccines were tested similarly: one group of randomly chosen volunteers received the real jab, and the other random half received a saline solution. However, in addition to the volunteers being unaware of the condition of the experiment in which they had been placed, the nurses administering the vaccination were unaware of its content and could not in any way unconsciously transmit that information. Such a study is called *double-blind*. This type of design is required to be able to validly conclude that a significant difference in COVID-19 prevalence between the two groups is due to the vaccination, in other words, that the vaccination works.

It will be clear to most readers that this experimental design setting is unrealistic when conducting empirical legal research. Law changes pertain to an entire country or union, and citizens are aware of the change. Also, in many settings, it would simply be impossible to randomise the intervention of interest. In the example above, we cannot randomise gender over court

cases: male or female citizens commit different crimes and have pertinent characteristics and behaviour that impact sentence length. Interviewing judges on whether they sentence male and female defendants differently is like asking them whether they act professionally in a breach of the constitution and is not likely to lead to valid responses.

In some instances, it is possible to investigate the impact of legally relevant phenomena using so-called *vignette studies*. In a vignette study, one presents a set of respondents with realistic but fictitious cases. For the example of gender effects on sentencing, such a vignette could be a police report or a court file in which a defendant has committed a violent crime. For a vignette, two versions of the court file are made: one in which the perpetrator is male and one in which the perpetrator is female. One distributes these different versions of the vignettes to judges and asks the judges what they believe an appropriate sentence would be. Now, the vignette is identical for the male and female defendants. No confounders are present that may explain differential sentencing: if a difference emerges between sentences for men and women, it can *only* be attributable to gender (and chance, obviously). By using testing, the likelihood of observing the sentence disparity by chance can be determined. If that likelihood is very small, we may conclude that gender indeed has an effect. A worked example can be found in Bijleveld et al. (2022). Van den Bos and Hulst (2016) discuss the possibilities and pitfalls of various kinds of experimental methods in empirical legal research.

6. Special topic: Systematic case law analysis

Systematic case law analysis is of particular relevance to ELS scholars. Hall and Wright (2008, p. 64) label it as a distinctly legal form of empiricism and state:

> Using this method, a scholar collects a set of document opinions on a particular subject, and systematically reads them, recording consistent features of each and drawing inferences about their use and meaning. This method comes naturally to legal scholars because it resembles the classic scholarly exercise of reading a collection of cases, finding common threads that link the opinions, and commenting on their significance.

In systematic case law analysis, one selects a sample (or an entire population) of opinions or court rulings, reads and codes the material, and searches to answer the research questions. Codes can be factual categories

such as *type of claim* or *gender of the litigant*, *chamber* or *background of the judge*, but they can also be derived from the material in the cases. Code selection using a large-scale systematic case law analysis is amply demonstrated in the well worked material by Wijntjens (2020).

Wijntjens' study investigated whether offering apologies to victims of harm by the party held liable for that harm induces the risk of being held liable in court. Offering apologies has been labelled as "legally dangerous" (Farmer, 2015, p. 244), as an apologetic statement may be admissible evidence at trial to establish liability or to prove some other element of an offence. Also, it has been noted that insurance companies may instruct the insured to be reticent in offering apologies and to speak only summarily and with great care on what happened, with mention made of lawyers even ordering their clients to remain silent (Cohen, 1999). Wijntjens (2020) studied to what extent the assumption that apologies might amount to an admission of liability in legal proceedings has an empirical basis in legal practice. The study employed systematic case law analysis, which differs from conventional legal analysis – in which issues are presented in one case or a small group of exceptional or weighty cases – in that it examines a large and representative group of cases to find overall patterns. As such, it aims to prove a claim not according to one author's rhetorical power but because the patterns that are found in case law have been uncovered through systematic and transparent empirical analysis of the rulings' content. Moreover, the data collection, data analysis and findings are reproducible.

The study selected court rulings from several databases with court rulings. Using keywords and by reading the rulings, Wijntjens arrived at a selection of 570 rulings in which apologies played a role. All texts were analysed and coded using a coding scheme that had qualitative and quantitative elements. First, the argumentative schemes that the judges used to arrive at their rulings were coded. Wijntjens coded whether apologies played a subordinate role, a conjunct role, or a decisive role in assessing the evidence on which the conclusion about the case would be based that the judge reached.

The results found that in very few rulings, apologies were decisive in the ruling. Out of all 570 coded and rulings analysed, only in seven judgments the court considered that the apologies of the person causing the damage as constituting an acknowledgement of liability. This amounts to 1.2%. Her findings clearly debunked the prevalent idea of the offering of apologies to be risky behaviour. Interestingly, the study also showed that withholding

apologies notably increased the risk of a negative outcome (Wijntjens, 2020).

7. Conclusion

This chapter could touch only very briefly on the various research methods available for empirical legal studies. While empirical research, and especially the more quantitative methods, may be relatively foreign to legal scholars, most are not very difficult to master. Experience teaches that both empirical and legal/doctrinal skills contribute to the production of sound empirical legal findings. Experience also teaches that empirical legal research is generally a journey of discovery, surprise and fun.

References

Bijleveld, C.C.J.H. (2023) *Research Methods for Empirical Legal Studies.* Den Haag: Eleven [Online]. Available at: https://elsacademy.nl/research-methods-for-empirical-legal-studies-an-introduction/ (Accessed: 9 February 2025).

Bijleveld, C.C.J.H., Blažević, M., Bociga Gelvez, D. and Buljubasic, M. (2022). 'Sanctioning Perpetrators of International Crimes: A Vignette Study'. *International Criminal Law Review*, 22, pp. 805-826.

Cane, P. and Kritzer, H. M. (2010) *The Oxford Handbook of Empirical Legal Research.* Oxford: Oxford University Press.

Cohen, J. R. 'Advising Clients to Apologize', *Southern California Law Review*, 72, pp. 1009-1069 [online]. Available at: https://ssrn.com/abstract=1612774 (Accessed: 14 January 2025).

Davies, G. (2020) 'The relationship between empirical legal studies and doctrinal legal research,' *Erasmus Law Review*, *2*, pp. 3–12.

Drost, E.A. (2011) Validity and Reliability in Social Science Research. *Education Research and Perspectives*, 38(1), pp. 105-124.

Dyevre, A. (2021) 'The promise and pitfall of automated text-scaling techniques for the analysis of jurisprudential change', *Artificial Intelligence and Law*, 29, pp. 239–269.

Farmer, C. (2015) 'Striking a Balance: A Proposed Amendment to the Federal Rules of Evidence Excluding Partial Apologies', *Belmont Law Review*, 2(243), pp. 243-267.

Hall, M. and Wright, R. (2008) 'Systematic content analysis of judicial opinions,' *California Law Review*, 96, pp. 63–122.

Haucke M., Hoekstra R. and van Ravenzwaaij D. (2021) *When numbers fail: do researchers agree on operationalization of published research?*, 8(9) [online]. Available at: https://doi.org/10.1098/rsos.191354 (Accessed: 14 January 2025).

Hutchinson, T. (2013) 'Doctrinal research: researching the jury' in Watkins, D. and Burton, M. (eds.) *Research methods in law*. London: Routledge, pp. 7-33.

Hutchinson, T. (2015) 'The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law', *Erasmus Law Review*, 3, pp. 130-138.

John, L.K., Loewenstein, G., Acquisti, A. and Vosgerau, J. (2018) When and why randomized response techniques (fail to) elicit the truth. *Organizational Behavior and Human Decision Processes, 148*, pp. 101–123.

Tracy, S. (2013) *Qualitative Research Methods*. Chichester: Wiley.

Tyler, T. (1990) *Why People Obey the Law*. New Haven, CT: Yale University Press.

Van Boom, W. H., Desmet, P. and Mascini, P. (2018) 'Empirical legal research. Charting the terrain' in Van Boom, W. H., Desmet, P. and Mascini, P. (eds.) *Empirical Legal Research in Action. Reflections on Methods and Their Applications*. Cheltenham: Edward Elgar, pp. 1– 22.

Van Gestel, R. and Micklitz, H.-W. (2011) 'Revitalising Doctrinal Legal Research in Europe: What About Methodology?', in Neergaard, U., Nielsen, R. and Roseberry, L. (eds.), *European Legal Method – Paradoxes and Revitalisation*. Kopenhagen: Djøf Publishing, pp. 25-73.

Van Velthoven, B.C.J. (2016) 'A Young Person's Guide to Empirical Legal Research. With Illustrations from the Field of Medical Malpractice', *Law and Method*, April 2016 [online]. Available at: https://doi.org/10.5553/REM/.000016 (Accessed: 14 January 2025).

Van den Bos, K. and Hulst, L. (2016) 'On Experiments in Empirical Legal Research', *Law and Method*, March 2016 [online]. Available at: https://doi.org/10.5553/REM/.000014 (Accessed: 14 January 2025).

Webley, L. (2010) 'Qualitative approaches to empirical legal research', in Cane, P. and Kritzer, H. M. (eds.) *The Oxford Handbook of Empirical Legal Research*. Oxford: Oxford University Press, pp. 927–950.

Wijntjens, L. (2020) *Als ik nu sorry zeg, beken ik dan schuld? Over het aanbieden van excuses in de civiele procedure en de medische tuchtprocedure*. The Hague: Boom Uitgevers.