UtrechtLawReview

Arthur Dyevre and Michal Ovádek, 'Experimental legal methods in the classroom' (2020) 16(1) Utrecht Law Review pp. 1–12.
DOI: https://doi.org/10.36633/ulr.557

ARTICLE

Experimental legal methods in the classroom

Arthur Dyevre* and Michal Ovádek†

As legal research and scholarship are increasingly turning to interdisciplinary approaches, the question arises as to how to introduce quantitative research techniques to a student population usually unfamiliar with empirical methods. We argue that classroom experiments form an effective — and, from the perspective of students, attractive — way to teach law students the logic of empirical inquiry. Many questions and controversies on and around adjudication and the impact of legal regulations hinge on matters of beliefs and behaviour which experimental methods are well-suited to investigate. Moreover, experimental legal research is fairly intuitive and does not require advanced statistical knowledge. Thanks to modern software tools, experiments can be conducted and analysed in the classroom without much prior technical knowledge. We provide basic guidance on how to undertake in-class experimental legal research and discuss examples of in-class experiments on gender effects, anchoring effects and neutrality bias.

Keywords: randomized experiments; legal methods; teaching; interdisciplinarity

1. Introduction

The legal academy is increasingly turning to interdisciplinarity.[1] While the empirical turn has been more pronounced in the United States[2] and Israel than in Europe, it is affecting both research and teaching.[3] Empirical legal studies have the potential to answer questions that are beyond the reach of traditional, doctrinal methods. These questions include the effective impact of legal rules on the behaviour of the human agents whose conduct they purport to regulate; the prediction of judicial outcomes; and the prejudices and biases of judges and other legal decision makers. In answering these questions, empirical research can generate insights that are useful to legal practitioners as well as legal reformers. We see the rise of empirical legal research as a positive development. Yet, as empirical legal research expands and diversifies, lawyers are increasingly expected, if not to be able to conduct empirical research themselves, at least to become intelligent consumers of empirical scholarship, both of the quantitative and qualitative variety.[4] This raises an educational challenge, for law students are usually unfamiliar with the concepts and techniques required to conduct and evaluate empirical work. Hence the question: how do we best impart empirical skills to this student population?

* Professor, Centre for Legal Theory and Empirical Jurisprudence, KU Leuven. Corresponding author. Email: arthur.dyevre@kuleuven.be. The authors gratefully acknowledge financial support from European Research Council Grant 638154 (EU-THORITY).

† PhD candidate, Centre for Legal Theory and Empirical Jurisprudence, KU Leuven. Email: michal.ovadek@kuleuven.be.

[1] Urska Sadl and Henrik Palmer Olsen, 'Can Quantitative Methods Complement Doctrinal Legal Studies? Using Citation Network and Corpus Linguistic Analysis to Understand International Courts' (2017) 30 Leiden Journal of International Law 327; Tom Ginsburg and Thomas J Miles, 'Empiricism and the Rising Incidence of Coauthorship in Law' (2011) University of Illinois Law Review 1785; Kees van den Bos and Liesbeth Hulst, 'On Experiments in Empirical Legal Research' [2016] Law and Method <https://www.bjutijdschriften.nl/tijdschrift/lawandmethod/2016/03/lawandmethod-D-15-00006> accessed 20 January 2019; Thomas J Miles and Cass R Sunstein, 'New Legal Realism, The' (2008) 75 University of Chicago Law Review 831; Peter Cane and Herbert Kritzer (eds), *The Oxford Handbook of Empirical Legal Research* (Oxford University Press 2010); András Jakab, Arthur Dyevre and Giulio Itzcovich (eds), *Comparative Constitutional Reasoning* (Cambridge University Press 2017).

[2] See Tracey E George, 'An Empirical Study of Empirical Legal Scholarship: The Top Law Schools' (2006) 81 Indiana Law Journal 141.

[3] Largely as a result of American influence on Israeli law schools, see Pnina Lahav, 'American Moment [s]: When, How, and Why Did Israeli Law Faculties Come to Resemble Elite US Law Schools?' (2009) 10 Theoretical Inquiries in Law 653.

[4] Christoph Engel, 'Legal Experiments: Mission Impossible?' (2013) Erasmus Law Lectures. One of the most cited impactful and most cited books ever penned by legal scholar leans extensively on experimental research, see Richard H Taler and Cass R Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (Penguin 2009).

We believe that empirical legal studies do not only enrich legal research but also legal teaching.[5] There are very few aspects, if any, of law and adjudication that empirical methods cannot help in some way to illuminate. Not only do we believe that this holds for all branches of law, but we also believe that students should be introduced to empirical methods early on in the legal curriculum. Knowledge of empirical methods, we believe, has benefits beyond the ability to read contributions in empirical legal journals. Indeed, they introduce students to basic concepts such as causality, variable measurement, data and probabilities, which are becoming increasingly useful to lawyers in the age of big data and legal tech.[6] Yet, having been teaching both large and small groups of law students with no prior knowledge of empirical research methods for several years, we are well aware of the educational challenge involved. Our experience suggests that in-class experiments represent a very effective, hands-on way to introduce law students to the logic of empirical research and to make them enthusiastic for interdisciplinary approaches to the study of law. First, the methodology of randomized (vignette) experiments is intuitive and very accessible, especially compared to observational studies, where the level of technical expertise required to overcome inferential problems can feel overwhelming. Second, thanks to modern online software tools, randomized controlled experiments can be conducted and even analysed in the classroom. The classroom effectively becomes a lab where students contribute actively to the production of research results. Third, experimental methods are relevant to many aspects of legal practice and judicial decision making. Experiments have shed light on a wide range of topics in litigation, adjudication and consumer protection, to mention but few.[7] Fourth, by showing how hypotheses can be formulated and tested against hard data, classroom experiments can generate great student enthusiasm. Not surprisingly, students are more likely to relate and engage with research that they have themselves contributed to produce. Experiments help cast light on the behavioural dimension of law and can make students aware of their own proclivities and cognitive biases. Finally, the results of in-class experiments can serve as basis for real academic research. They thus represent a very hands-on manner to bring teaching and research together.

Once students have become familiar with experimental designs, it becomes easier to introduce them to other qualitative and quantitative methods: qualitative interviews, survey designs, field experiments, statistical methods for observational data, etc. In-class experiments give them a feel for the concepts and principles that form the basic building blocks of empirical research: variables, hypothesis, causality, external and internal validity and probabilities.

In this paper we do not only explain how to conduct in-class legal experiments but we also point out some pitfalls to avoid and discuss examples of in-class legal experiments we have conducted ourselves. In addition, we discuss how students respond to in-class experiments. As we shall see, our experience suggests that students respond enthusiastically and perceive the value of this kind of research for the study of law as well as for legal practice.

2. How to conduct in-class legal experiments

Experiments come in various guises. They can be conducted in real-world social settings. They are then called 'field experiments'. But they can also be conducted in a lab-like environment. Experiments can take a game-like form or operate with vignettes – short descriptions of a hypothetical scenario. Participants can be students, judges or legal counsels. Here, we concentrate on randomized vignette experiments, because they are the most relevant to legal research and can be run in the classroom.[8]

[5] For a spirited plea for empirical approaches in legal research see Arthur Dyevre, Wessel Wijtvliet and Nicolas Lampach, 'The Future of European Legal Scholarship: Empirical Jurisprudence' (2019) 26 Maastricht Journal of European and Comparative Law 348.

[6] See Daniel Martin Katz, 'MIT School of Law – A Perspective on Legal Education in the 21st Century, The' (2014) 2014 University of Illinois Law Review 1431.

[7] Chris Guthrie, Jeffrey J Rachlinski and Andrew J Wistrich, 'Blinking on the Bench: How Judges Decide Cases' (2007) 93 Cornell Law Review 1; Andrew J Wistrich and Jeffrey J Rachlinski, 'How Lawyers' Intuitions Prolong Litigation' (2012) 86 S. Cal. L. Rev. 571; Andrew J Wistrich, Jeffrey J Rachlinski and Chris Guthrie, 'Heart versus Head: Do Judges Follow the Law of Follow Their Feelings' (2014) 93 Texas Law Review 855. See also example below.

[8] The relevance of vignettes for legal research is most easily seen in light of the prevalence of the case-based method of teaching. A vignette in fact often comprises a case scenario, sometimes even directly adapted from real-life situations or court decisions. We also recognize the somewhat different nature of in-class experiments from laboratory conditions. In the latter the researcher has full control over every detail of the experimental setting. In-class experiments are to some extent constrained by course requirements and student experience.

2.1 Basics of experimental research: Treatment and randomization

People otherwise unfamiliar with social science research often have a rough idea of the basic steps required to conduct experiments from medical research, where randomized controlled experiments constitute a prominent research method. Experiments invariably involve dividing participants in separate groups. In medical research, this usually means in treatment and control groups. In drug trials, for example, the treatment group will receive the drug while the control group will receive a placebo. In law and the social sciences, participants are also divided in separate groups, except that they more typically receive vignettes rather than drugs and placebos. A vignette is a description of a hypothetical situation – e.g. a hypothetical legal dispute – to which the subject is asked to respond. Subjects are presented with slightly different versions of the same vignette depending on the group to which they have been assigned. The difference between the versions of the vignette is what the experiment is designed to measure and defines the treatment and control groups. In the examples discussed below, the two (sometimes three) groups of students are presented with a vignette that is in all respects identical save for the element that the research aims to test – the defendant's motion requesting that the case be transferred to a small-claim court, the gender of the victim and the reasons justifying the choice of a particular legal basis.

It is essential for the validity of experimental findings that participants be assigned randomly to the distinct treatments, i.e. the versions of the vignette. Law students, like human subjects in general, are not identical. They can differ in multiple respects: age, grades, legal skills, socioeconomic background, political opinions, motivation, etc. Without randomization these (observed or unobserved) personal attributes may affect group assignment and distort the results. By ensuring that each participant has the same probability to be assigned to one group or the other, randomization equalizes these personal characteristics across the groups. The likelihood of encountering left-leaning, strongly motivated students or conservative, unmotivated students will be the same across groups. This obviates the need to control for these characteristics at the analysis stage, thereby making the number-crunching aspect of experiments considerably easier. Random assignment is straightforward to implement in the survey software we recommend for running in-class legal experiments.

Randomization may also be used to control other aspects of experiments. A typical pitfall in experimental design is overlooking how in the presence of multiple questions or complex vignettes ordering can affect responses. For example, if subjects are asked two questions relating to a vignette, it may be wise to randomize the order of the questions. Otherwise it might be difficult to exclude the possibility that the order in which the questions are presented is not affecting their answer, for example because students may improve their understanding of the material after answering the first question (learning effect).

2.2 How to evaluate the results of in-class experiments: Average treatment effect

Once the subjects' responses have been collected, assessing the results of a randomized vignette experiment is less difficult than those unfamiliar with empirical methods might expect, especially if the experiment is conducted using online survey software. Basically, assessing the results involves comparing the pattern of responses between groups (although sometimes we may be interested in *within-* rather *between*-group variation). What we want to do is to compare the groups' average response. This is called, in more technical terms, the average treatment effect (ATE). If the students' response is a choice within a continuous range of possibilities (e.g. amount of damages to grant, length of prison sentence, etc.), a box-and-whisker plot provides, as we shall demonstrate below, a very intuitive tool to visualize and assess differences between and within groups – although formal tests using analysis of variance (ANOVA) or ordinary least square (OLS) regression can also help and are usually required for the results to be published.

When the response is not continuous but binary (e.g. guilty/not guilty), the analysis can be compared to assessing the probability of coin flips coming up heads or tails. Although, unlike coin tosses, the probability of someone being found guilty is rarely exactly 50%, we can use similar statistical techniques to understand probabilities of binary outcomes. They can be visually represented by likelihood functions and their statistical significance can be tested using, for example, bivariate logistic regression or Pearson's Chi-squared test. Behind the nitty gritty technical and mathematical details peculiar to these techniques lies the same underlying intuition as for any formal statistical test: we use a theoretical, probabilistic distribution as a benchmark to determine whether what we observe — that is, the difference between groups — is more than just a random fluctuation. If the difference between groups deviates significantly from the probabilistic distribution, we can make a good case that the difference in treatment is systematically affecting behaviour. The most often used threshold for considering results statistically

significant is a p value lower than 5%,[9] but other thresholds can be justified in the context of the research design.[10]

Although running these tests requires a basic knowledge of statistics, their technical implementation is nowadays relatively straightforward thanks to statistical software. One of the most user-friendly, free and open-source options is JASP but analysis of experiments can be easily carried out in all popular statistical software such as R, SPSS, Stata or Matlab. Plenty of tutorials can be found online on how to implement the abovementioned statistical tests (and others) in any of these software packages.

2.3 Practical tips

Writing good, effective vignettes is more difficult than it may seem at first sight. Small differences in the language of a vignette can have a large impact on the outcome of an experiment.[11] For instructors who have never run experiments, we recommend that they start by replicating existing ones (like the first two reported below) before setting out to design their own vignettes. Replication, apart from being increasingly appreciated scientifically,[12] is an effective way to get familiar with experimental research methods. What is more, as Van den Bos and Hulst notice, replications can bring up issues which might not have been sufficiently addressed in the original work, thereby contributing new knowledge.[13]

While completing the experiments students should not know what hypothesis is being tested or which group they are being assigned to. Accordingly, the instructor/investigator should refrain from informing them until the experiment is completed. Ideally, the instructor should also ignore how the students have been assigned until their response have been collected. This is because knowledge of the experiment's specific goal or the researcher's inadvertent behaviour may unconsciously influence participants and undermine the validity of the results.

In-class experiments can be pen-and-paper or online-based. Over the former the latter has the advantage that it automatically converts the students' responses into machine-readable spreadsheets which can be fed into statistical software, thus significantly reducing the time needed to analyse the results.[14] In addition, using online tools ensure that experiments are effectively double-blinded. Group assignment is not known to the student, nor to the teacher/investigator. In that way, there is no risk that they inadvertently bias the results. Online survey platforms, such as Qualtrics or SurveyMonkey, offer a user-friendly environment to implement randomized experiments. The treatment and control version of a vignette can be written as separate survey questions with students then randomly assigned to one or the other.

In general, the more participants in an experiment, the better. A larger sample size means that if the hypothesized effect is really present, the researcher is more likely to detect it. Conversely, if no effect is found, we can be fairly certain that the hypothesis can be rejected in light of the data. In other words, a larger sample size confers more statistical power. The concept of statistical power can be expressed as the probability of a particular test rejecting the null hypothesis if a given alternative hypothesis is true. Power analysis can be conducted in advance in order to get a sense of what sample size is required to observe the hypothesized effect.[15] This said, an in-class experiment does not have to produce an effect in order to be interesting and to illustrate the principles of empirical legal

[9] Another way to describe the 5% p value threshold is by stating that the researcher accepts a long-run Type I error rate of 5%. Type I errors are false positives: situations when the researcher wrongly rejects the null hypothesis.

[10] Daniel Lakens et al, 'Justify Your Alpha' (2018) 2 Nature Human Behaviour 168.

[11] Herman Aguinis and Kyle J Bradley, 'Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies' (2014) 17 Organizational Research Methods 351; Peter M Steiner, Christiane Atzmüller and Dan Su, 'Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap' (2016) 7 Journal of Methods and Measurement in the Social Sciences 52; Spencer C Evans and others, 'Vignette Methodologies for Studying Clinicians' Decision-Making: Validity, Utility, and Application in ICD-11 Field Studies' (2015) 15 International journal of clinical and health psychology 160.

[12] Marcus R Munafò, Brian A Nosek, Dorothy Bishop, Katherine Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis, 'A Manifesto for Reproducible Science' (2017) 1 Nature Human Behaviour 1.

[13] Kees van den Bos and Liesbeth Hulst, 'On Experiments in Empirical Legal Research' [2016] Law and Method <https://www.bjutijdschriften.nl/tijdschrift/lawandmethod/2016/03/lawandmethod-D-15-00006> accessed 20 January 2019.

[14] Occasionally though, pen-and-paper experiments can offer the teacher more control over the experimental conditions in class.

[15] For example, an experiment in which we expect a certain frame to make the treated group to award on average 25% more damages to victims of car accidents than the control group would require 199 participants in each group, given an accepted Type I error rate of 5% and 80% power (which are the default values in most experiments).

research. Moreover, an experiment can be run in multiple classes in order to achieve greater statistical power.[16]

When it comes to ethical requirements, in-class experiments must abide by the same principles underlying regular laboratory or field experiments, such as avoidance of harm. In practice, legal experiments conducted in class normally pose fewer ethical issues, as they are non-invasive and harmless. Nonetheless, it is paramount that informed consent is always obtained from subjects and that data protection rules and other university-wide ethical requirements are fully respected.[17] Depending on applicable standards and experimental design, prior approval from the university's ethics committee might be necessary.

3. Examples of in-class experiments

In this Section we present examples of vignette experiments we have conducted with law students in the classroom. The first replicates a prior study. The other two are original experiments designed by the two authors.

3.1 Anchoring and legal argumentation

Anchoring, also known as 'focalism', is a cognitive bias characterized by the tendency to rely too heavily on initial information, even straightforwardly absurd or irrelevant, when making decisions. The effect of anchoring is illustrated by a famous study conducted by Daniel Kahneman and Amos Tversky in which participants observed a roulette predetermined to stop on either 65 or 10 and were then asked to guess the proportion of African countries in the United Nations. Participants who had observed the wheel (randomly) stopping on 10 came up with lower estimates (25% on average) than those whose wheel stopped on 65 (45% on average).[18]

It is easy to see how anchoring may influence decision making in the legal context. For example, the defender in settlement negotiations may try to influence the outcome by making a very low initial offer so as to create a low anchor. The plaintiff may want to do the exact opposite. Similarly, parties to a tort dispute may seek to influence the judge with either low or high anchor dependent on whether they want to minimize their liability or maximize the amount of damages awarded.

Adapting an experiment by Chris Guthrie, Jeffrey Rachlinski and Andrew Wistrich,[19] we studied the impact of anchoring in a hypothetical tort case in which students acted as judges. Our hypothesis was that students exposed to a low anchor in the form of a motion from the defendant's legal counsel would be less generous in the amount of damages granted than those not exposed to that anchor. The participating students were randomly assigned to two groups. The control group—the no anchor group—was presented with the following vignette:

> Suppose that you are presiding over a personal injury lawsuit. The defendant is a major company in the package delivery business. The plaintiff was badly injured after being struck by one of the defendant's trucks when its brakes failed at a traffic light. Subsequent investigations revealed that the braking system on the truck was faulty, and that the truck had not been properly maintained by the defendant. The plaintiff was hospitalized for several months, and has been in a wheelchair ever since, unable to use his legs. He had been earning a good living as a free-lance electrician and had built up a steady base of loyal customers. The plaintiff has requested damages for lost wages, hospitalization, and pain and suffering, but has not specified an amount. How much would you award to the plaintiff in compensatory damages?[20]

The treatment group—the anchor group—was presented with the exact same vignette except for the following mention which was added at the end of the vignette: 'The defendant has moved for dismissal of the

[16] When computing the ATE for responses from multiple classes, it is recommended that the researcher/instructor add a dummy variable for each class to its statistical model so as to control for unobserved class-specific effects. (If students are not randomly assigned to classes but choose which class to attend, this may entail selection effects.)

[17] The ethics of experiments should in itself form one of the in-class points of discussion. In order to preserve the integrity of the experimental design, a practical solution is to do a debriefing and obtain consent *ex post*.

[18] Amos Tversky and Daniel Kahneman, 'Judgment Under Uncertainty: Heuristics and Biases' (1974) 185 Science 1124, 1128.

[19] Guthrie, Rachlinski and Wistrich (n 7) 19.

[20] As some anonymous reviewers observed, there are a number of additional facts and elements — such as the age and marital status of the plaintiff, the defendant's history regarding compliance with safety standards, etc. — which a real-world judge would normally ask to know before coming to a decision. These limitations, though, stem from the original design of the experiment by Guthrie et al.

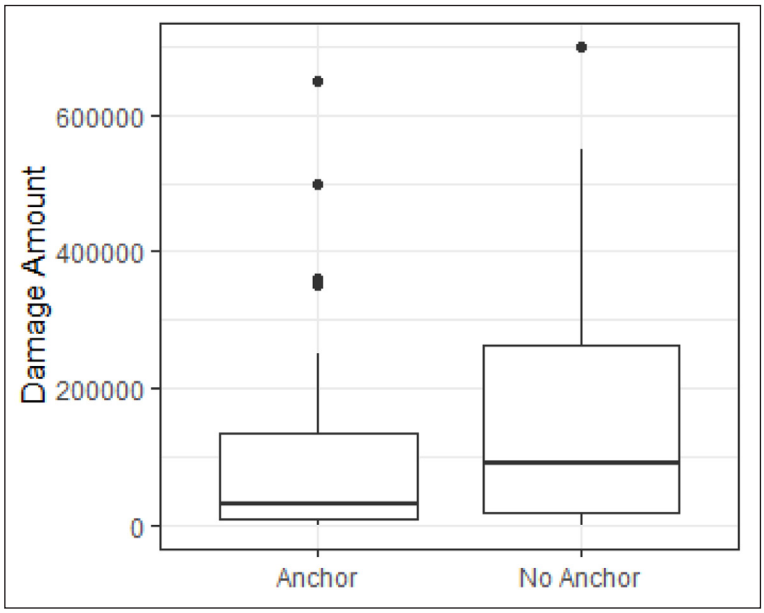

Figure 1: Box-and-whisker plot of student responses to the vignette with and without the low anchor.

case, arguing that cantonal courts alone have jurisdiction over claims below 25,000 EUR'. This defendant's motion served as our anchor. The experiment was implemented as an online survey using Qualtrics. 166 students from an undergraduate course taught at a Belgian law school participated in this experiment. The experiment was double blind. The students did not know that they were randomly assigned to different groups, nor did they know the hypothesis being tested. Using automated randomization also ensured that we could not observe assignment during the experiment itself. This is important as subconscious cues may otherwise influence participants. Of course, students were informed of the hypothesis and results in a subsequent debriefing in accordance with the terms of the consent form.

Analysing the results, we found that some students had given extreme values (one granted several trillions Euros). We removed these extreme responses to facilitate the analysis (notably data visualization), but this is not a required step—in our case this does not change the substantive conclusion from the experiment. **Figure 1** illustrates the difference between the treatment and control group with the help of box-and-whisker plot.

A box-and-whisker plot divides the values of the responses in quartiles. The values below the box represent the first quartile, those in the lower half of the box the second quartile, those in the upper half the third quartile and the values above the box represent the fourth quartile. The dots represent outliers—values far from the mean value. The horizontal bar within the box represents the median. This means that 50% of students awarded damages below this value and 50% above. Here we see that, compared to the control group, the median in the treatment group—the group exposed to the anchor—is substantively lower. The distribution of values is generally more compressed and closer to the median. This suggests a systematic difference between the two groups consistent with our anchoring hypothesis. This is confirmed by regression analysis. On average, students not exposed to the anchor awarded 60 776 Euros more to the plaintiff – a difference unlikely to result from chance.[21] In the study that inspired our tort experiment subjects were real-world judges, who were found to display strong sensitivity to anchoring.[22] Our results indicate that law students are prone to the same bias.

The experiment shows how lawyers can exploit the cognitive biases of judges to win in the courtroom. A claim—such as an absurdly high damage claim or an absurdly low damage assessment—may influence the judge even when the judge knows it is wrong. In the classroom, this provides a good opportunity to discuss the psychology of legal decision makers along with the role of non-legal factors in adjudication.

[21] Our statistical software reported a p-value of 0.02, which means that if we were to run that experiment 100 times, we would observe a difference of that magnitude just by chance no more than two times.

[22] Guthrie, Rachlinski and Wistrich (n 7).

3.2 Gender in criminal law

In another experiment we undertook to explore the effect of gender on sentencing. What we wanted to test is whether male and female law students view sexual assault cases differently and, in particular, whether male students may be more lenient towards male perpetrators. Research on the effect of sex on judging suggests that, other things being equal (that is, after eliminating the effect of other factors such as ideology), female and male judges do not behave differently except in areas where there is experiential and informational asymmetry such as gender discrimination.[23] Sexual assault may plausibly fall under this exception—men and women are likely to have differing experiences and fears regarding sexual assault.

To be sure, because in sexual assault cases women are typically the victims and men the perpetrators, constructing a good control is difficult. How can we tell whether male students are more lenient because the facts of the case are perceived to be different or because they are genuinely less harsh towards male offenders? To untangle these effects, we designed an experiment with three conditions: (1) an assault case, without a sexual element, in which the victim is a woman; (2) the same assault case but with a male victim; and (3) the same assault case but with a sexual element and a female victim (the perpetrator is male in all three conditions). Our hypothesis was that there would no significant within-group difference between male and female participants in Group 1 and 2, but that male students would be more lenient than female students in Group 3. In other words, the gender barrier in experience and perceptions of sexual violence should only be manifest in Group 3.

Students in Group 1 were presented with the following vignette:

> Amanda M., a 23 year old bank clerk, is about to get in her car when she is brutally assaulted by Jerome D. Jerome D. punches her in the face, severely damaging her lower jaw, before robbing her of her valuables – including cash money, jewellery, a smartphone and a tablet – and running away. Soon thereafter Jerome D. is arrested and charged with physical assault and robbery. As the assault was recorded by video surveillance and Amanda M.'s belongings were retrieved inside his home, Jerome's guilt is attested. As a judge, you must decide on the length of Jerome D.'s prison sentence. By law, physical assault is punished with 8-day to 2-year imprisonment and robbery with 6 months to 10 years. Convicts may serve consecutive sentences.
>
> Please indicate the LENGTH IN MONTHS (not in years) of the prison sentence Jerome D. should serve.

In Group 2 students were presented with the exact same vignette, save for the fact the victim was a he (Robert M.)[24] rather than a she. For Group (3) the vignette added a sexual element:

> Amanda M., a 23 year old bank clerk, is about to get in her car when she is brutally assaulted by Jerome D. After groping her breasts and forcing a kiss on her, he punches her in the face, damaging her lower jaw, before robbing her of her valuables – including cash money, jewellery, a smartphone and a tablet – and running away. Jerome D. is soon arrested and charged with physical assault, sexual assault and robbery. As the assault was recorded by video surveillance and Amanda M.'s belongings were retrieved inside his home, his guilt is attested. As a judge, you must decide on the length of Jerome D.'s prison sentence. By law, physical assault is punished with 8-day to 2-year imprisonment, sexual assault with 6 months to 5 years imprisonment and robbery with 6 months to 10 years imprisonment. Convicts may serve consecutive sentences.
>
> Please indicate the LENGTH IN MONTHS (not years) of the sentence Jerome D. should serve.

Two hundred thirty-four students from the same undergraduate course at the same Belgian law school participated in this experiment. As shown in **Figure 2**, the results do not lend support to our hypothesis. True, the median prison sentence in Group 3 is lower for male than for female students whereas the median for male is higher than for female in Group 1 and identical to that of female students in Group 2. However,

[23] Christina L Boyd, Lee Epstein and Andrew D Martin, 'Untangling the Causal Effects of Sex on Judging' (2010) 54 American Journal of Political Science 389. See also Erik J. Girvan, Grace Deason, and Eugene Borgida, 'The generalizability of gender bias: Testing the effects of contextual, explicit, and implicit sexism on labor arbitration decisions' (2015) 39 Law and Human Behavior 525.

[24] The authors were informed by reviewers that 'Robert M.' is the name of a famous Dutch perpetrator. However, we believe that the case was not sufficiently salient in Belgium to affect participants in the experiments. This assumption is consistent with the fact that results are nearly identical for Group 1 and Group 2.

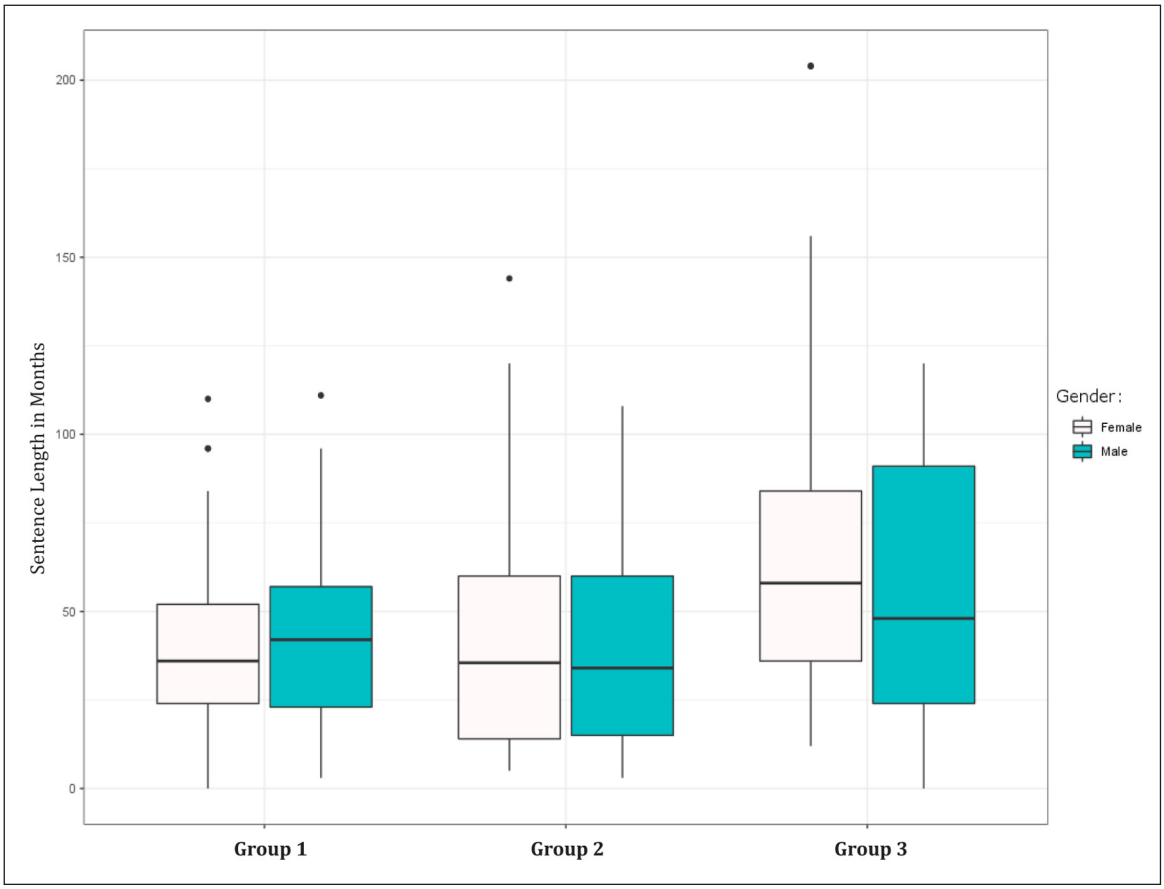

Figure 2: Months of imprisonment imposed by male and female students after reading a scenario in which the victim of assault was a woman (Group 1), the victim of assault was a man (Group 2), or the victim of a sexual assault was a woman (Group 3).

values for male students in Group 3 exhibit a large variance, which means their values are more dispersed (less compressed around the median than those of female participants).

A formal test (regression analysis) confirms that while there is a significant between-group difference – average sentence length is significantly longer in Group 3 – the within-group difference between male and female students in Group 3 is not statistically significant.[25]

Interestingly, we also invited students to report their ideology on a 10-point scale. Unlike gender, this factor turned out to have a significant effect on sentence length. Holding the facts of the case constant (by statistically removing the effect of scenario), more right-wing students handed out prison sentences that, on average, were significantly longer.[26] On average, a one-point increment on the 10-point ideological scale is associated with an additional 3.8 month imprisonment.

As Belgian legal education tends to emphasise black-letter law and seldom attends to the role of extraneous factors in adjudication, the results provided a good opportunity to engage students' perceptions and assumptions about judicial decision making. As part of the experiment, we asked students to rank the factors that have the strongest influence on judicial decision making, 68.6% ranked legal rules first; 7.5% ranked gender first; 6.8% the arguments presented by the parties; 5.4% the position of the other courts. Only 5.1% ranked ideology first. The prevailing, formalist picture of judging thus offered an interesting contrast to the results of the experiment.

3.3 Neutrality bias

Legal systems and cultures rest on various fictions and normative aspirations. Two of the most important ones, certainly in the European context, are objectivity and neutrality. Lawyers typically regard or act as if they regard the law and legal decision-making to be objective and neutral (non-partisan). As a result, it is

[25] The p-value is 0.000125, which translates in approximately 1 chance in 10000 to obtain such a result by chance.

[26] The p-value for this result is 0.003200.

reasonable to expect lawyers—especially less experienced law students—to have at least unconscious attachment to prima facie objective and neutral information when they do their work.

An original experiment was conducted to find out how would-be lawyers react when information – irrelevant from a legal or doctrinal standpoint – is presented as 'political'.[27] The term 'political' conveys for many the notions of interest and partisanship which to some extent stand in opposition to the values of objectivity and neutrality. Given that lawyers, like other decision-makers, rely on heuristics when engaging in legal reasoning,[28] we should be able to uncover the impact of legal arguments being 'tainted' as political regardless of their analytical irrelevance.

The experiment was carried out at a Belgian law school in the context of a class on European Union (EU) law. The class specifically addressed the concept of legal basis in EU legislation. Eighty students were presented with two draft laws and asked to select for each an appropriate legal basis (a provision of the EU Treaties) from two given options, following the doctrine of the European Court of Justice (ECJ) that students had been told to study.[29] The two draft laws (A and B) differed in ambiguity: whereas one was intended to be perfectly poised between the two competing legal bases, the other called straightforwardly for one legal basis of the two given, in light of the ECJ doctrine and the wording of the provisions. Thus, using the analogy of a coin flip, draft law A was designed in such a way as to essentially resemble a fair coin – one where the odds of heads and tails are 50–50 – whereas draft law B was very much biased towards one outcome.[30]

The treatment consisted of a sentence in which the subjects were informed that one of the legal bases was being favoured by an EU institution 'for political reasons'. The control assignment contained no such information. It was expected that the 'political' condition would push treated subjects towards the 'untarnished' legal basis at statistically significant levels.

In **Table 1** we can see that, indeed, as hypothesized, the treatment group was more likely to select the 'non-political' legal basis than the control group. The ATE tells us how much more likely – on average – the treated group of students was to select the 'non-political' legal basis compared to the control group. Consistent with the hypothesis, the ATE has a positive sign for both scenarios, but the effect size varies. On aggregate, the treatment group was 17% more likely to select the 'non-political' legal basis. The greater ambiguity of scenario A arguably diminished the potency of the effect, as both legal options were relatively plausible. The more clear-cut scenario B generated an effect twice as large, showing that the hypothesized treatment could erode confidence in even the most obvious legal solution.

Effect sizes alone do not tell us, however, whether the observed differences between the groups can be considered very surprising from a statistical standpoint; they could be merely results of random variation. As a consequence, we need to test for statistical significance, using, for example, bivariate logistic regression or Pearson's Chi-squared test. Both tests show that when aggregating the observations from the two scenarios, the results are statistically significant.[31] The same holds true for legislation B taken separately but not for legislation A. This difference can be attributed to the different degrees of ambiguity – the effect

Table 1: Number of participants (n) and successes (x) per group (control/treatment) and scenario (A/B). Success x is defined as the choice of the 'non-political' legal basis under treatment conditions. The average treatment effect *ATE* is calculated as $(\frac{x_{treatment}}{n_{treatment}} - \frac{x_{control}}{n_{control}}) * 100$. The hypothesis predicted the *ATE* to have a positive sign, meaning that political treatment makes the choice of the 'non-political' legal basis more likely.

Draft law	$n_{control}$	$n_{treatment}$	$x_{control}$	$x_{treatment}$	*ATE*
A	45	42	18	22	12.4%
B	42	45	2	13	24.1%
A + B	87	87	20	35	17.2%

[27] While the present article was under review, a detailed description of this experiment was published in Michal Ovádek, 'The apolitical lawyer: experimental evidence of a framing effect' (2019) 48 European Journal of Law and Economics 385.

[28] Gerd Gigerenzer and Christoph Engel (eds) *Heuristics and the Law* (MIT Press 2006).

[29] The choice had to be motivated with legal reasoning. In this way the experiment engaged at the same time the more traditional skill of legal writing. This experiment therefore also illustrates that measurable quantities are not a sine qua non of this type of research, despite the overall thrust towards law and economics in the literature. Motivating choices in writing stimulated subsequent in-class discussion among students.

[30] As can be seen in **Table 1**, the 50–50 ratio was not actually achieved. In the control group, the ratio was closer to 40–60.

[31] The *p*-value is less than 0.01 level in both the parametric (logistic regression) and the non-parametric (Pearson's Chi-squared) test.

is more difficult to detect under conditions of greater ambiguity. The experiment therefore turned out to be underpowered for scenario A; a larger sample size would arguably yield a statistically significant result at the 5% level even under the conditions of greater ambiguity. Nonetheless, we can overall conclude that the hypothesized aversion to the 'political' on the part of would-be lawyers was supported by the experimental data.

What we can also glean from this experiment is the didactically useful demonstration of different degrees of ambiguity (or indeterminacy) of the law. All legal practitioners and scholars would probably agree that depending on both facts and law, a given situation can be more or less indeterminate from a legal perspective. Experiments can make this aspect of the law more tangible for students; in the present case, we can clearly observe that the probability of one legal basis being selected over another differs depending on the legal and factual scenario. These differences between probabilities of different outcomes of a legal analysis offer a fresh way of conveying not only the idea that the law might not always provide a single correct answer to a legal problem but that some legal outcomes are nonetheless more plausible than others.

4. Students' perceptions of in-class experiments

How do students respond to in-class experiments? We surveyed law students who participated in in-class experiments. We asked them whether experiments are helpful for comprehension; whether they would endorse their broader use in the law school curriculum; whether in-class legal experiments are useful to understand law in practice and whether they enhance the learning experience. We reached out to students from two different law courses taught during the 2018–19 academic year. The results are reported in **Table 2**.

While the sample is small, the response rate was high (80%). The vast majority of surveyed students (over 80%) acknowledged at least some positive effect of in-class experiments on their understanding of the materials and empirical research methods discussed in the class. Students also appear to be enthusiastic about this teaching approach. Seventy-eight percent of respondents either agreed or strongly agreed that in-class experiments make the learning experience more enjoyable. Contrary to perceptions of empirical insights as irrelevant for the practice of law, most students regarded experiments as enhancing their understanding of how law is applied in practice. Seventy-one percent of respondents were of the opinion that in-class legal experiments can 'definitely' or 'probably' help them in better understanding the practical application of law. Given this pattern of responses, it is unsurprising that a large majority of students surveyed (82%) would endorse broader use of in-class experiments in legal education.

Table 2: Student perception of in-class experiments (n = 28). We asked in total 35 students (response rate = 80%) from two different law classes in which in-class experiments were used.

Comprehension and learning experience						
Strongly Agree	**Agree**	**Somewhat Agree**	**Neither agree nor Disagree**	**Somewhat Disagree**	**Disagree**	**Strongly Disagree**
Q1: Do you think in-class experiments facilitate comprehension of the class materials?						
21%	50%	18%	7%	0%	4%	0%
Q2: Do you think in-class experiments are helpful for understanding empirical research?						
32%	39%	21%	4%	4%	0%	0%
Q3: Do you think in-class experiments make the learning experience more enjoyable?						
39%	39%	14%	4%	4%	0%	0%

Educational use and relevance for legal practice				
Definitely yes	**Probably yes**	**Might or might not**	**Probably not**	**Definitely not**
Q4: Would you endorse broader use of in-class experiments as a teaching technique in law schools?				
32%	50%	11%	7%	0%
Q5: Do you think in-class experiments help you to better understand the practical application of the law?				
32%	39%	21%	4%	4%

5. Conclusion

In this article we sought to promote the use of in-class experiments as an intuitive tool to impart empirical insights to law students as well as a viable bridge between teaching and research. Experiments are typically more straightforward to analyse than observational quantitative data and with the help of online software are easy to implement in the classroom. These advantages make in-class experiments a well-suited tool to introduce law students to empirical legal methods. We explained the building blocks of a randomized vignette experiment and gave some practical tips to avoid common but easily preventable mistakes. We presented examples of experiments which, aside from being easily replicable, help give concrete contours to the methodology and its application in the legal context. Moreover, a small-scale survey among students exposed to in-class experiments reveals a positive impact on perceived understanding class materials and the overall learning experience. Students appear to be enthusiastic about in-class experiments, which they view as relevant to legal training and legal practice.

Readers and instructors interested in running in-class experiments can look up the expanding literature for further illustrations and inspiration. The work of Jeffrey Rachlinski, Andrew Wistrich, Chris Guthrie and Christoph Engel—who count among the most prominent names of experimental legal research—presents fascinating explorations of the effects of biases and heuristics on legal decision-making.[32] Experiments have didactic as well as scientific value that legal teachers and researchers would be imprudent not to take advantage of.[33] While a certain degree of investment in understanding experimental designs and key concepts is necessary, legal experiments need not become lectures on psychology or economics. In-class experiments can concern relatively specific biases manifest only in a given area of law (such as criminal justice) which pertain directly to the legal questions canvassed in the class materials. Moreover, greater student engagement can foster better overall retention of class materials. Going forward and with an eye on the legal education reform debate, we believe that creating incentives for students to learn experimental methods and apply them in-class would enlarge both the students' skillset and their understanding of law in practice.

Competing Interests

The authors have no competing interests to declare.

[32] From a vast bibliography: Jeffrey J Rachlinski, Andrew J Wistrich, and Chris Guthrie, 'Probability, Probable Cause, and the Hindsight Bias' (2011) 8 Journal of Empirical Legal Studies 72; Andrew J Wistrich and Jeffrey J Rachlinski, 'How Lawyers' Intuitions Prolong Litigation' (2013) 86 Southern California Law Review 571; Chris Guthrie, Jeffrey J Rachlinski, and Andrew J Wistrich, 'Inside the Judicial Mind' (2001) 86 Cornell Law Review 777; Christoph Engel and Lilia Zhurakhovska, 'You Are In Charge – Experimentally Testing the Motivating Power of Holding a Judicial Office' (2017) 46 The Journal of Legal Studies 1; Christoph Engel and Michael Kurschilgen, 'The Coevolution of Behavior and Normative Expectations: An Experiment' (2013) 15 *American Law and Economics Review* 578; Christoph Engel et al, 'First impressions are more important than early intervention: Qualifying broken windows theory in the lab' (2014) 37 International Review of Law and Economics 126; Theodore Eisenberg and Christoph Engel, 'Assuring civil damages adequately deter: A public good experiment' (2014) 11 Journal of Empirical Legal Studies 301.

[33] While this essay is not the proper place to address the comparative strengths and weaknesses of qualitative, experimental and quantitative-observational approaches, we believe that it is in their ability to identify cognitive mechanisms that lies the main comparative advantage of experimental designs.

How to cite this article: Arthur Dyevre and Michal Ovádek, 'Experimental legal methods in the classroom' (2020) 16(1) Utrecht Law Review pp. 1–12. DOI: https://doi.org/10.36633/ulr.557

Published: 26 May 2020

Utrecht Law Review is a peer-reviewed open access journal published by Utrecht University School of Law.

OPEN ACCESS