



Empirical Coding of Judicial Opinions: A Practical Guide for Doctrinal Scholars

Ngozi Menon^{1*}, Aurelie Zhaksybekov², Freya Kowalski³

¹Department of Law, Transylvanian University Centre for Law, Netherlands

²Department of Law, Dalmatian University Centre for Law, Taiwan

³Department of Law, Vistula School of Law, Kyrgyzstan

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijls.2026.v1.i1.pp14-19](https://doi.org/10.62715/ijls.2026.v1.i1.pp14-19)

Received: 03 November 2025

Accepted: 23 December 2025

Published: 08 March 2026

Abstract

Doctrinal scholars read cases closely and report what they find, but they seldom say how the cases were chosen or how another reader could check the account. Systematic coding of judicial opinions is a way of doing the same work transparently: a defined set of opinions is read against a written protocol, the features of interest are recorded in a consistent form, and the reliability of the recording is tested. The technique is well established in political science and communication research and has been used in legal scholarship for several decades, yet many lawyers regard it as requiring statistical training they lack. This article offers a practical guide written for the doctrinal scholar. It sets out a seven-stage procedure covering the framing of a question that counting can answer, the definition of the population of opinions, the drafting of a codebook, piloting, coder training and reliability testing, full coding, and analysis and reporting. For each stage it explains the decisions to be made and the errors that most often undermine published studies. The procedure is assembled from the methodological literature and is offered as guidance; it is not applied to a set of cases here, and the article says what the method cannot show.

Keywords: *Content analysis, Judicial opinions, Empirical legal research, Codebook, Intercoder reliability, Doctrinal method*

© 2026 Ngozi Menon et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

A scholar who writes that courts "generally" apply a test in a certain way, or that a factor is "rarely decisive", is making an empirical claim. The usual support for it is a string of cases the author regards as representative. The reader cannot tell how many cases were read, whether contrary ones were set aside, or whether a second lawyer reading the same material would have described it in the same terms. Doctrinal research has many strengths, among them a trained sensitivity to legal reasoning that no counting exercise can replace (Hutchinson and Duncan, 2012), but verifiability is not one of them.

Systematic content analysis addresses that weakness. In the form described for lawyers by Hall and Wright (2008), the researcher selects cases by a stated rule, codes each one for defined features, and then

* Corresponding author: Ngozi Menon, Department of Law, Transylvanian University Centre for Law, Netherlands. E-mail: ngozi.menon@transylvanian.edu

© 2026 Ngozi Menon et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

counts and compares. The result does not displace interpretation. It disciplines it, by forcing the scholar to state in advance what is being looked for and by exposing the whole population of cases instead of the memorable few.

This article is a guide to the method for scholars whose training is doctrinal. It proposes a procedure in seven stages, drawn from the methodological literature in law and in the social sciences. The procedure is not demonstrated on a body of case law here, and no findings are reported; the aim is to explain what to do, in what order, and why.

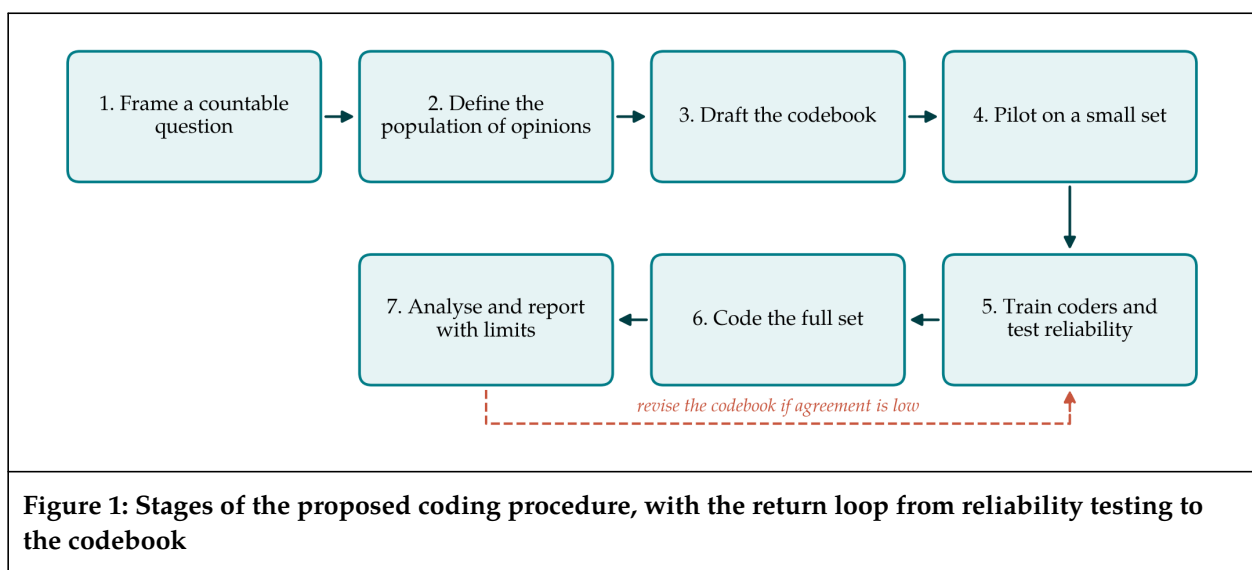
2. Background: What Coding Can and Cannot Do

Content analysis is a research technique for making replicable and valid inferences from texts to the contexts of their use (Krippendorff, 2018). Applied to judgments, it has a respectable record. Thompson (1991) coded a large body of decisions on piercing the corporate veil and showed that several propositions repeated in the textbooks did not describe what courts did. Sisk et al. (1998) coded district court opinions on the constitutionality of the federal Sentencing Guidelines to study influences on judicial reasoning. Beebe (2008) coded reported fair use opinions under the United States Copyright Act and examined how the four statutory factors related to one another and to the result. Each study answered a question that a doctrinal scholar would recognise as doctrinal.

The method also has well-known limits. An opinion is a justification written after the decision and addressed to particular audiences; it is evidence of the reasons a court was willing to give, not a transparent record of why it decided (Edwards and Livermore, 2009). Reducing a judgment to categories discards nuance, and categories borrowed from political science can misdescribe law, a criticism made forcefully of the issue codes in the standard database of United States Supreme Court decisions (Shapiro, 2009). Coders who know the hypothesis may code toward it (Harvey and Woodruff, 2013). Hall and Wright conclude that the method suits the study of what opinions say and how often, and is weaker as a means of explaining or predicting outcomes (Hall and Wright, 2008).

3. The Proposed Procedure

The seven stages are shown in Figure 1. They are presented in sequence, but the early stages are iterative: piloting almost always sends the researcher back to the codebook, and poor reliability should do so again.



3.1. Stage 1: Frame a question that counting can answer

Good coding projects begin with a claim that could turn out to be false. "How do courts treat proportionality?" is a topic. "Courts that cite the structured proportionality test apply all of its limbs in most cases" is a claim with observable implications, which is what Epstein and King (2002) ask of any

empirical legal argument. The scholar should write down the claim, the features of an opinion that would count for and against it, and any rival explanation worth recording. Questions about frequency, co-occurrence and change over time suit the method. Questions about what the law ought to be do not, though coded results may inform them.

3.2. Stage 2: Define the population of opinions

The next task is to say which opinions the claim is about and how they will be found: the courts, the period, the database and the exact search terms or citation rule. Where the population is manageable, all of it should be coded; where it is not, a random sample drawn from a complete list is preferable to a selection by prominence (Epstein and Martin, 2014). Two selection effects need to be acknowledged in every project. Reported opinions are not a random draw from decisions, and their differences from unreported ones can be large (Siegelman and Donohue, 1990). Decided cases are not a random draw from disputes either, because parties settle the cases whose outcome is predictable (Priest and Klein, 1984). Conclusions should be stated as conclusions about the population actually coded.

3.3. Stage 3: Draft the codebook

The codebook is the central document of the project. For each variable it gives a name, a definition, the permitted values, a decision rule for hard cases and at least one worked example. Categories must be exhaustive and mutually exclusive, which usually means including "other" and "unclear" values and watching how often they are used (Neuendorf, 2017). Variables differ in how much judgement they demand, and Table 1 groups the kinds a doctrinal project typically needs. A sound rule is to code what the opinion states and to keep inference to a minimum: whether a court cited a given authority is easier to code reliably than whether it "really" relied on it. Outcome variables should be coded separately from reasoning variables, ideally at a different time, so that knowledge of the result does not colour the reading of the reasons.

Table 1: Types of variable in a codebook for judicial opinions, with typical difficulties and safeguards			
Type of variable	Example	Typical difficulty	Safeguard
Identifying	Court, date, citation, panel size	Clerical error	Take from database fields; spot-check
Party and procedure	Type of claimant; stage of proceedings	Mixed or multiple parties	Rule for the dominant category
Outcome	Claim allowed, dismissed or allowed in part	Partial success; remittal	Define success from one party's side
Authority cited	Whether a named precedent or provision is cited	String citations; indirect reference	Code express citation only
Doctrinal step	Whether each limb of a test is addressed	Limbs merged or implicit	Worked examples; an "implicit" value
Reasoning style	Reliance on purpose, text or policy	High inference; coder disagreement	Two coders; report agreement

3.4. Stage 4: Pilot on a small set

Before any coding counts, the researcher applies the draft codebook to a small number of opinions taken from the population but excluded from the later reliability test. Piloting shows which definitions are ambiguous, which categories are never used and which facts cannot be found in the text. The codebook is revised and the changes are logged. Scholars working alone should still pilot, and should ask a colleague to code a few opinions cold, since the author of a codebook is the person least able to see its gaps.

3.5. Stage 5: Train coders and test reliability

Reliability is the extent to which different coders, working independently from the same instructions, record the same values. It is tested by having at least two people code the same subsample without conferring and comparing the results variable by variable. Simple percentage agreement flatters, because some agreement arises by chance; the usual corrections are the kappa coefficient of [Cohen \(1960\)](#) for two coders and nominal categories and the alpha coefficient of Krippendorff, which accommodates several coders, missing values and different levels of measurement ([Hayes and Krippendorff, 2007](#)). The benchmarks most often quoted are those of [Landis and Koch \(1977\)](#), who described kappa between 0.61 and 0.80 as substantial agreement, and of [Krippendorff \(2018\)](#), who advises relying on variables with alpha of at least 0.80 and drawing only tentative conclusions from those between 0.667 and 0.80. These are conventions and should be reported as such.

Where a variable falls short, the remedy is to clarify the definition, retrain and test again on fresh opinions, or to drop the variable. It is not to discuss disagreements until they disappear and then report the agreed figures. The coefficient, the size of the subsample and the procedure should appear in the published article, an expectation long standard in communication research ([Lombard et al., 2002](#)). Coders should not be told the hypothesis where that can be avoided.

3.6. Stage 6: Code the full set

Once reliability is acceptable the codebook is frozen and the full population is coded, with values entered in a spreadsheet or database in which each row is an opinion and each column a variable. The unit of analysis should be settled beforehand: the case, the majority opinion, each separate opinion, or each issue within an opinion. Long projects should include a second reliability check partway through, because coders drift. Every departure from the codebook, and every opinion excluded after retrieval, is recorded with reasons.

3.7. Stage 7: Analyse and report with limits

For most doctrinal purposes the analysis is modest: frequencies, cross-tabulations and trends over time, presented in tables a lawyer can read. Regression has its place when several factors must be considered together, but it requires care and usually a collaborator, and it adds nothing if the underlying coding is unsound. The report should describe the search, the population, the codebook and the reliability results in enough detail for another scholar to repeat the work, and the codebook and coded data should be made available, which is the replication standard [Epstein and King \(2002\)](#) urged on law reviews. Counts should then be read alongside the opinions themselves. The strongest studies quote and discuss cases to show what a category looks like in practice.

4. Fitting the Method to Doctrinal Work

Three adjustments make the method workable for a scholar without a research team. The first is scale. A project of a few dozen to a few hundred opinions with ten or fifteen variables is feasible for one author and a second coder for the reliability subsample, and is often enough to test a doctrinal claim. The second is division of labour: the author's legal expertise goes into the codebook, and the coding itself is designed to need as little of that expertise as possible. The third is integration. Coding should be planned as one part of a doctrinal article, supplying the descriptive base on which the interpretive and normative argument then builds. A scholar who has coded every appellate decision on a provision is also unusually well placed to write about it in the traditional way.

Some features of legal material call for particular care. Appellate judgments often contain several opinions, and a variable such as "the court applied the test" has no single value where the judges divided on the point; the codebook must say whether the majority, the lead judgment or each opinion is coded. Cases that raise several issues may need one row per issue. Reversals and later history matter if the claim is about the law as it stands and not about judicial behaviour at the time. Jurisdictions also differ in how much of the reasoning is written down: a tradition of short, impersonal judgments offers fewer codable features than one of long individual opinions, and a comparative project must take care that it is not

measuring drafting style. None of these problems is peculiar to coding. Ordinary doctrinal reading meets them as well, and the difference is that a codebook makes the chosen solution visible.

5. Limitations and What Remains to Be Shown

The guide is a proposal and has not been validated here by application to a set of opinions. Its thresholds for reliability are conventions imported from other disciplines, and it is an open question whether they are well calibrated for legal texts, in which reasonable lawyers disagree about meaning. The procedure assumes opinions that are published and reasoned; it has less to offer for jurisdictions where judgments are brief or selectively reported. It leaves aside automated text analysis, which may reduce the labour of coding but raises validity questions of its own. And it cannot cure the basic limitation that an opinion records stated reasons. A coded study describes judicial writing with a precision that ordinary reading lacks. It does not reveal what judges thought.

6. Conclusion

Coding judicial opinions asks the doctrinal scholar to do openly what is otherwise done privately: choose cases by a rule, read them against stated criteria and show that the reading is repeatable. The seven stages described here require patience more than statistics. Used with an awareness of selection effects and of the gap between reasons given and reasons held, the method strengthens doctrinal claims about what courts do, and leaves the scholar free to argue about what they should do.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Beebe, B. (2008). [An empirical study of U.S. copyright fair use opinions, 1978-2005](#). *University of Pennsylvania Law Review*, 156(3): 549-624.
- Cohen, J. (1960). [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1): 37-46. <https://doi.org/10.1177/001316446002000104>
- Edwards, H.T. and Livermore, M.A. (2009). [Pitfalls of empirical studies that attempt to understand the factors affecting appellate decisionmaking](#). *Duke Law Journal*, 58(8): 1895-1989.
- Epstein, L. and King, G. (2002). [The rules of inference](#). *University of Chicago Law Review*, 69(1): 1-133. <https://doi.org/10.2307/1600349>
- Epstein, L. and Martin, A.D. (2014). [An Introduction to Empirical Legal Research](#). Oxford University Press, Oxford.
- Hall, M.A. and Wright, R.F. (2008). [Systematic content analysis of judicial opinions](#). *California Law Review*, 96(1): 63-122.
- Harvey, A. and Woodruff, M.J. (2013). [Confirmation bias in the United States Supreme Court Judicial Database](#). *Journal of Law, Economics, and Organization*, 29(2): 414-460. <https://doi.org/10.1093/jleo/ewr003>
- Hayes, A.F. and Krippendorff, K. (2007). [Answering the call for a standard reliability measure for coding data](#). *Communication Methods and Measures*, 1(1): 77-89. <https://doi.org/10.1080/19312450709336664>
- Hutchinson, T. and Duncan, N. (2012). [Defining and describing what we do: Doctrinal legal research](#). *Deakin Law Review*, 17(1): 83-119.
- Krippendorff, K. (2018). [Content Analysis: An Introduction to Its Methodology](#), 4th edn. Sage, Thousand Oaks.

- Landis, J.R. and Koch, G.G. (1977). [The measurement of observer agreement for categorical data.](#) *Biometrics*, 33(1): 159-174. <https://doi.org/10.2307/2529310>
- Lombard, M., Snyder-Duch, J. and Bracken, C.C. (2002). [Content analysis in mass communication: Assessment and reporting of intercoder reliability.](#) *Human Communication Research*, 28(4): 587-604. <https://doi.org/10.1111/j.1468-2958.2002.tb00826.x>
- Neuendorf, K.A. (2017). [The Content Analysis Guidebook, 2nd edn.](#) Sage, Thousand Oaks.
- Priest, G.L. and Klein, B. (1984). [The selection of disputes for litigation.](#) *Journal of Legal Studies*, 13(1): 1-55. <https://doi.org/10.1086/467732>
- Shapiro, C. (2009). [Coding complexity: Bringing law to the empirical analysis of the Supreme Court.](#) *Hastings Law Journal*, 60(3): 477-540.
- Siegelman, P. and Donohue, J.J. (1990). [Studying the iceberg from its tip: A comparison of published and unpublished employment discrimination cases.](#) *Law & Society Review*, 24(5): 1133-1170. <https://doi.org/10.2307/3053664>
- Sisk, G.C., Heise, M. and Morriss, A.P. (1998). [Charting the influences on the judicial mind: An empirical study of judicial reasoning.](#) *New York University Law Review*, 73(5): 1377-1500.
- Thompson, R.B. (1991). [Piercing the corporate veil: An empirical study.](#) *Cornell Law Review*, 76(5): 1036-1074.