UNIVERSITY OF COPENHAGEN
FACULTY OF SOCIAL SCIENCES



Like Cases, Unlike Outcomes

Arbitrariness in Administrative Discretion, Decisions, and Delivery

Karl-Emil Svenningsen Bendtsen

PhD dissertation

Karl-Emil Bendtsen

Like Cases, Unlike Outcomes

Arbitrariness in Administrative Discretion, Decisions, and Delivery

A dissertation presented for the degree of

Doctor of Philosophy

31 August 2026



Summary

A basic promise of the rule of law is that like cases should be treated alike. What citizens receive should follow from the law and the facts of the case, not from whoever administers it. However, every law must be applied to particular people and implemented by particular bureaucrats. Across four papers, this dissertation asks the broad question: how often, and why, do like cases receive unlike outcomes?

Discretion for frontline bureaucrats is therefore central to the dissertation. Prior research has mostly examined whether bureaucrats, on average, drift from whatever an appropriate decision might be; whether the system is biased; whether the average bureaucrat treats different citizens differently. Fewer studies have examined whether bureaucrats differ from one another on the same case; whether the system has variance; whether different bureaucrats treat like cases differently. Whereas the first approach holds the bureaucrat fixed and looks at variation across cases, then at averages, the second holds the case fixed and varies the bureaucrat, then looks at the spread.

This dissertation does the second. As such, it looks at the extent to which decisions are arbitrary. Could a given case have received a different outcome, had it simply been allocated a different bureaucrat? This is a counterfactual, and it can be measured without a benchmark of what the correct decision of a case would be.

For this, we need random allocation of bureaucrats. To get this data, the dissertation in Paper 1 repurposes the underlying data of 95 leniency instrumental variable studies covering more than 25 million cases by almost 100,000 bureaucrats. These could be cases like foster care referrals, incarceration, or patent granting. In the median setting, the allocation decides at least 13% of cases, and more than 29% in a quarter of the settings. Across the middle half of settings, two randomly drawn bureaucrats disagree on 3.3% to 12.9% of cases at minimum; if a case has a correct decision, disagreement implies one bureaucrat was wrong, so in most settings at least 1.7% to 6.5% of decisions are errors. These numbers only capture the arbitrariness that comes from differences in general leniency. Holding that constant, the dissertation furthermore re-analyses 13 conjoint experiments, in which actual bureaucrats are forced to choose between helping one of two different cases; they disagree 40% of the time.

How, then, can we reduce arbitrariness and make like cases get like outcomes? The dissertation draws on two of its papers to theorise how to reduce arbitrariness. It suggests strengthening support for bureaucrats as they make judgments, for example with AI, and making laws and rules more specific to align decision-making.

Like cases can also receive unlike outcomes because of factors on the citizen side. The last paper concludes with a meta-analysis of interventions that seek to strengthen take-up of social benefits. It finds that across a large range of countries and settings, informing citizens of their eligibility and helping them comply with the process is effective in increasing take-up.

Acknowledgements

The worst part of doing a PhD has been the occasional loneliness. I am therefore very grateful for the people who have contributed to making the process of writing this dissertation be less lonely and making my life very entertaining and intellectually enjoyable. Whereas academia can often be cynical, for me, I have mostly experienced the opposite; many people have looked out for me, helped and guided me or just made every day be more fun.

The main paper of the dissertation measures how much people are dependent on the bureaucrat that is assigned to handle their case. When I got a spot in the PhD program, I myself was assigned to my supervisor Mogens Jin Pedersen. I am really grateful for that. This dissertation would have looked a lot different, been much worse, or never existed if it had not been for Mogens. Throughout the years, Mogens' rigour has benefitted my chaotic approach to academic work and academic life, although his patience with me probably has not always been easy to maintain. I continue to aspire to be as disciplined and thorough as he is. Discussing research with Mogens always helps me to cut to the chase in my own thoughts and ideas (although we disagree on the academic value of quirky research). Without Mogens, I would not have made it.

I also want to thank Asmus Leth Olsen. I started working with Asmus as a student research assistant prior to me entering the PhD program, and over the years he is probably the one who has helped and influenced me the most in developing my taste in good research. Asmus has taught me a lot on how to see through fads and fluff—which at times can be especially prevalent in our field of research—instead to focus on the actual quality of evidence and the big questions of public administration. He continually inspires me to think bigger.

My time at CSS would not have been quite as fun without some of my very talented and very sweet colleagues from the PhD program. Thank you to my good friend and office mate, Benedikte Grundtvig Huber, for bringing extravaganza and decadence to the office and for always doing the most in making every day more social, gossipy and overall fun. I hope to share an office with Benedikte for the rest of my working life.

Thank you to my good friend, co-author and co-organizer of the hit work event series "Vinstitut for Smagskundskab", Frederik Klaaborg Kjøller. Kjøller is always contributing to making the workplace less mundane. He is an extremely good colleague, and he is always looking out for and taking care of people as he has helped many like me by opening up and talking about the harder sides of being a PhD student. A good work day has often been one where I have been at his office, receiving puns and talking about wine.

Also, thank you to my long-time friend Christoffer Hentzer Dausgaard for our many discussions of big ideas about political science research. Besides being one of the nicest people in academia in its entirety, he is probably also one of the smartest. Discussing with him, it often feels like he is one step ahead of my own thoughts. I aspire to be able to understand a problem so fast and so clearly as him. Christoffer always makes things seem simple and almost obvious, in that way, he has made me understand the movie *Amadeus* and how Salieri must have felt.

From both the Public Administration and Public Policy group and the PhD group, I want to thank Rasmus Fredslund Stenderup and Josephine Ebba Thorup Arnfred. To Rasmus, a guy who is at his loudest and giddiest when he talks about public administration questions and research, I am grateful for all of our discussions of PA theory, cycling and manager games. To Josephine, suspiciously more quiet about public administration questions and research, I am grateful for her often reminding me that I was doing alright although I have whined a lot to her. She is also a very good, straightforward researcher and talking to her always makes me happy.

Thank you to Olivia Levinsen for her dry wit, Daniel Cruz Doggenweiler for his laconic humour and the boozed discussion over food cultures, Benedikte Raft for being good at taking initiative and for being so sweet, Frederik Jerris for banter and for gossip, Matilde Bro Hansen for her amusing directness, Esther Ejrnæs Bastholm for being so happy and optimistic, Gabriela Štvrtňová for alerting me to some political theory for the introduction, and Ingrid Helene Brandt for generally being very sweet. Also thank you to the other younger scholars whom I have shared many presentations and discussions with including Marie Kongsted Møller, Marie Heldgaard Seestedt, Søren Damsbo-Svendsen, Nijat Eldarov, Vicente Alves, Petra Mathilde Jørs, Mariia Vladymyrova, Debora Del Piano, Lily (Yi-Hui) Lin, Kristina Aleksandrovna Pedersen, Karen Philippa Larsen, Anna Lund Sørensen, Maevia Laureen Griffiths.

The PhD group and my enjoyment within owe a lot to probably the greatest PhD administrator in the world, Lise Sonnichsen who always does the most for the group.

The PhD school leader, Jacob Hariri, also deserves a big thank you. Jacob is one of the coolest and most charismatic people in academia, I am sure; I hope to be as great a showman as him one day. Besides that, he is also really smart and he is the one who has given me the best comments on my papers as they were in the making.

I also want to thank Frederik Hjorth. Frederik has helped me not only during the past three years but also during my time at the master's level and the bachelor's. He was always willing to engage with me and help me in a manner with huge respect, creating the illusion of almost feeling like being at the same level. Before the PhD, Frederik helped encourage me to take on harder methodological tasks than I thought I was capable of, from which I learned a lot and which gave me academic confidence. I would probably not have thought about doing a PhD had it not been for the informal guidance from him.

In a similar vein, I want to thank Greg Eady. He has been both my co-supervisor on the PhD dissertation and he was the supervisor of my master's thesis. I have often consulted Greg with questions of methods, in order to get his highest praise of a short "yeah, you can do that" or "cool". Conferring with Greg has probably been one of the most efficient steps in doing research for me as he very quickly and clearly sees the problems and the possibilities of any idea, helping me to avoid going in circles.

I also want to thank Lene Holm Pedersen for taking a great interest in my work and for pushing me to ask different questions to try to understand the bureaucracy and public policy. I co-wrote a paper with Lene outside this dissertation which taught me a lot on writing a good paper. Similarly, I also want to thank my other co-authors on projects outside the dissertation, Jacob Strabo, Paul van Leeuwen, Asmus Leth Olsen and Frederik Kjøller.

For the two co-authors on the papers within the dissertation, I also owe a huge thanks.

Thank you Jonas Krogh Madsen and Florian Keppeler.

I also want to thank Elizabeth Linos and The People Lab at Harvard Kennedy School for hosting me in the spring of 2025. It was a huge learning experience to see how research can be elevated when a team functions so well. There were many people in the People Lab and at HKS who deserve my thanks; Tanya Dall, Caroline Meserve, Alec Wall, Heidi Wallace, Noelia Negri, Max Spohn, Brenda Sciepura, Nadia Bell, Audrey Feldman, Jessica Lasky-Fink, and Karalyn Lacey.

My time in Cambridge, MA, would also not have been the same without the great Marc Sabatier Hvidkjær showing me around and introducing me to the Department of Government at Harvard. I always enjoy talking with him, as Marc le Bavard can talk for hours with such great passion and enthusiasm about political science as a discipline, a skill he also uses in being the global hub for gossip on the people in the global political science community; getting these updates has been very entertaining during the American summer and winter breaks where I have had the pleasure of sharing an office with him.

I would also like to thank Paul Hünermund, Scott Cunningham, Benjamin Carl Krag Egerod and Florian Hollenbach for great courses in advanced causal inference, and Martin Vinæs Larsen for generally being very cool and for lending me some data for a project I did not end up finishing for the dissertation. For fun in the lunch room, I also want to thank Dean Cooper-Cunningham and Lene Hansen.

Thank you to the people who have given me comments throughout the years at various conferences and seminars. A special thank you goes to Roman Senninger, Jesper Asring Hansen and Martin Bækgaard, and Frederik Hjorth, Christoffer Dausgaard and Jacob Hariri.

I also want to thank the committee, Asmus Leth Olsen, Simon Calmar Andersen and Maria Cucciniello.

From outside academia, I would like to thank my friend Marius Overgaard for making the art for the dissertation, and I would like to thank my family for having been so patient with me during the past years.

Finally, I especially want to thank the love of my life, Asta, for having supported me throughout this process. Thank you for always believing in me, for keeping my inflated ego inflated, for helping me not to be boring, and for generally making my life warm, loving and fun.

Contents

Dissertation Frame **7**

1 Introduction **7**
1.1 The bureaucrat and the citizen . . . 8
1.2 The concept of arbitrariness . . . 10
1.3 Structure of the dissertation . . . 12

2 Discretion and arbitrariness **15**
2.1 The original problem of discretion and the Finer-Friedrich debate . . . 16
2.2 How street-level bureaucrats actually are . . . 19
What Lipsky saw . . . 19
What Lipsky did not see . . . 20
2.3 Are there right answers? . . . 22
2.4 The moral street-level bureaucrat . . . 24
2.5 Does the discretion literature measure arbitrariness? . . . 26
Discrimination of citizens . . . 28
Bureaucrat types . . . 29
Footnotes to Lipsky . . . 29
2.6 Have other literatures answered the question? . . . 31
Dispersion in adjudication . . . 31
2.7 Conclusion . . . 32

3 What Is Arbitrariness? **35**
3.1 The counterfactual object . . . 35
3.2 Benchmarks and wrong decisions . . . 36
Deviations from the benchmark . . . 38
3.3 The general case . . . 40
3.4 Multifaceted bureaucrats . . . 42
3.5 Is this not just discrimination? . . . 44
Additional analysis: Comparing discrimination and leniency . . . 46
3.6 Consistently wrong: tails, thresholds, and aggregation . . . 47
The bias–variance trade-off . . . 47
Thresholds . . . 48
Panels . . . 50
3.7 Conclusion . . . 51

4 How to Get Like Cases Treated Alike **55**
4.1 On the citizen's role . . . 55
4.2 Two sides of simplification . . . 57
4.3 The official's rule . . . 57
4.4 The judgment around the rule . . . 59
4.5 The citizen's process . . . 60
4.6 Conclusion . . . 61

5 Key findings **65**
5.1 Paper 1: Arbitrary Bureaucratic Decisions . . . 65
5.2 Paper 2: AI's Role in Enhancing Consistency . . . 67
5.3 Paper 3: Longer Laws, Better Service? . . . 67
5.4 Paper 4: Increasing Take-up of Social Benefits . . . 68
5.5 Implications and future research . . . 70

Research Articles **73**

6 Arbitrary Bureaucratic Decisions **75**
6.1 Introduction . . . 76
6.2 Theory . . . 78
The Problem of Arbitrary Discretion . . . 78
Administrative errors and arbitrariness . . . 79
6.3 Study 1: How much arbitrariness is there? Re-analysis of leniency-IV studies . 81
Sample construction . . . 82
Estimands . . . 83
Estimation . . . 86
6.4 Results . . . 87
How many decisions are arbitrary? . . . 87
For the lucky and unlucky ones . . . 91
What could explain arbitrariness? Discretion and preferences . . . 93
6.5 Study 2: Re-analysis and variance decomposition of conjoint experiments with bureaucrats . . . 96
Sample construction . . . 99
Results . . . 99
6.6 Discussion and conclusion . . . 103

7 Noise in Administrative Decision-Making: Public Administration's Unseen Error and AI's Role in Enhancing Personnel Selection Consistency **107**
7.1 Introduction . . . 108
Research Focus . . . 109
Contribution to Public Administration . . . 110
7.2 The Concept of Noise . . . 111
The Problem of Noise . . . 112
Noise and Bias . . . 112
The Challenge to Public Values . . . 114
7.3 Towards More Consistent Decision-Making . . . 115
Algorithms and AI . . . 116
Human-AI Collaboration . . . 117
7.4 Data and Design . . . 120
Data Source: Prior Experimental Studies . . . 121
Current Study: Data and Operationalization . . . 121
Analytical Strategies . . . 122
7.5 Results . . . 124
Behavioral Responsiveness to AI Advice . . . 124
Descriptive Overview of Rating Distributions . . . 124
Analysis of Noise Reduction . . . 124
7.6 Robustness and Auxiliary Analyses . . . 129
7.7 Discussion and Conclusion . . . 129
Importance for PA . . . 130

Methodological Limitations . . . 131
Perspectives for Future Research . . . 132
Implications for Practice . . . 133

8 Longer Laws, Better Service? Legislative Complexity and Bureaucratic Quality 137
8.1 Introduction . . . 138
8.2 Theory . . . 140
8.3 Empirical context and data . . . 143
Measuring complexity and discretion in legal texts . . . 145
8.4 Research design and empirical strategy . . . 147
Conditioning on complaints . . . 148
8.5 Results . . . 149
Unpacking the change in length . . . 150
Endogeneity and reactions to errors . . . 151
8.6 Conclusion and Discussion . . . 152

9 Increasing Take-up of Social Benefits: A Meta-analysis of Field Experiments 155
9.1 Introduction . . . 156
9.2 The study of take-up . . . 157
Interventions and administrative burdens . . . 158
9.3 Methods . . . 159
Coding of application stage and administrative burdens . . . 160
Estimation . . . 161
9.4 Results . . . 163
Descriptive statistics . . . 163
Estimated average effects . . . 164
Decreasing learning demands or decreasing compliance demands . . . 166
Publication Bias . . . 168
9.5 Discussion and conclusion . . . 171
9.6 Data availability . . . 173
9.7 Acknowledgements . . . 173

References 174

Danish summary 194

A Reanalysis of the Israeli small-claims courts 197

B Sources for Danish simplification initiatives 201

C External applications of the SEWP 207

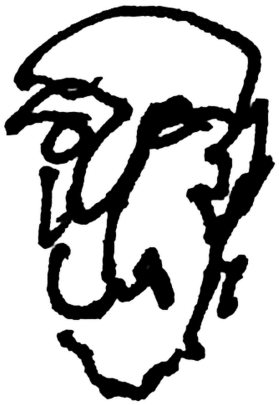
D Paper 1 Appendices 209
D.1 Complier share among recipients: the P10–P90 contrast . . . 209
D.2 How the variance decomposition follows from the model . . . 210
D.3 Variance decomposition model specification . . . 210
D.4 AMCE validation . . . 211
D.5 Disagreement and partisanship in Liu (2024) . . . 212
D.6 Geographic distribution of studies . . . 213
D.7 Additional leniency-IV results . . . 214
D.8 IV assumptions in the underlying studies . . . 219
D.9 How selected is the leniency-IV sample? . . . 221

D.10 Additional conjoint results . . . 228
D.11 Subgroup comparisons . . . 231
D.12 Estimation details . . . 232
D.13 Getting the numbers without replication data . . . 233
D.14 Where each number comes from . . . 235
D.15 Raw vs. residualized instruments . . . 236
D.16 Web of Science Search Query for Leniency-IV Studies . . . 236
D.17 Re-studied systems . . . 236
D.18 Caseload correlation: mechanical relationship check . . . 237
D.19 Is the rise at rare margins mechanical? . . . 238
D.20 Leniency IV studies, measures, and sources . . . 240

E Paper 2 Appendices **251**
E.1 Theoretical Elaborations . . . 251
I. Distinction: Bias and Noise . . . 251
II. Connection: Noise and Legitimacy . . . 251
III. Concerns: AI Deployment in Government . . . 252
IV. Preference: Specialization over Ensembling . . . 252
E.2 Sample Characteristics . . . 253
E.3 Derivation of the 'Specialization-Ensemble Weighting Point' (SEWP) . . . 254
E.4 Results of Levene's Test . . . 254
E.5 Robustness Analysis by Candidate Profiles . . . 254
E.6 Robustness Analysis Using the Full Sample . . . 257
E.7 Methodological Limitations . . . 259

F Paper 3 Appendices **263**
F.1 Further descriptives . . . 263
F.2 Panel balance . . . 266
F.3 Is there collider bias? . . . 268
F.4 Results measured with logit . . . 269
F.5 Results for other text-based measures . . . 271
F.6 Precedent rulings and the timing of legal change . . . 271
F.7 Parliamentary debates: do legislators invoke the appeals system? . . . 273
F.8 The consultation record . . . 274
F.9 Classifying the content of paragraph changes . . . 276

G Paper 4 Appendices **279**
G.1 PRISMA . . . 279
G.2 Further descriptives . . . 296
G.3 Search string . . . 298
G.4 Publication bias plots and tables . . . 299
Andrews & Kasy models . . . 300
G.5 Regression tables . . . 301
Overall average effect . . . 301
Risk ratio . . . 302
G.6 Interaction with variable for extra steps . . . 303
Differences in learning demands interventions? . . . 303
Difference across control base-level . . . 305
G.7 Further notes on estimation . . . 307









Chapter 1

Introduction

One of the most basic principles of the rule of law is the principle of equal treatment: like cases should be treated alike. What a citizen receives follows from the law rather than from whoever administers it. Whereas the formal policies settle what the bureaucracy should do, for whom, and by which facts and conditions, the actual decisions and delivery of those policies are made by the people involved in the process. The idea is that officials implement and administer those policies faithfully, *sine ira et studio*, as the Weberian ideal of bureaucracy dictates (Weber, 1978, p. 225), and that citizens then receive the treatment they are entitled to. If this arrangement works as it should, nothing else should affect the outcome: not the official who opens the file, and not the citizen's skill in working the process.

This highlights the fundamental, democratic problem in policy implementation. Fuller (1964) argued that congruence between announced rules and their administration is one of the basic conditions for legality. However, all policies need to be implemented and administered, and depending on how much can be specified and determined ex ante, all policies and laws leave room for interpretation and judgement, require verification, and must ultimately be applied to individual cases by specific decision-makers. Administration is the last link in the chain of democratic control, but it is one nobody votes on. As such, policies are not necessarily what they are decided to be. As Lipsky (2010) argues, policies are in reality what ends up at the citizen's end, i.e., after interacting with the bureaucracy. It matters less what the law says, if the citizen never receives the treatment the law promises.

In bureaucracies, we typically only discover errors by looking twice. Maybe a citizen appeals a decision, or a manager reopens a file, or an oversight body audits whether the right rules were applied. If the complaint is upheld or the audit uncovers that the correct procedure was not followed, an error can be identified and attached to the particular decision. But does an administrative error need to be checked in order for us to know that it happened? If two citizens are completely alike on paper, and they are treated differently by the bureaucracy, then we can infer that something went wrong. We can imagine two bureaucrats handling a case: they receive the same information, apply the same rules and standards, and reach different decisions. While we cannot determine who was right and who was wrong, if there is a correct answer, we can still infer that one of them made an error. If there are multiple correct answers, it still shows that the citizen's treatment depended on the official assigned and the case was thus still decided by arbitrary factors outside the case. We can also think of a citizen not being able to complete various application processes for social benefits, where the compliance requirements function less as a verification device and more like a second eligibility test on whether the citizen is sufficiently knowledgeable and skilled. Errors like these are not readily visible. We can rarely observe these ordinary, everyday injustices. As Shklar (1990, p. 112) writes, "Much is silent, forgotten, or locked away," while "the grinding injustices of daily life are endured." What goes wrong may

never become a complaint, an audit finding, or a reversed decision.

What makes bureaucratic institutions different from charity or private service provision is that they have to adhere to the same standards for everyone. Provision of services in bureaucracies is made to be "compulsory, collective, and largely nondiscretionary" (Goodin, 1988, pp. 10–12). Citizens are entitled as a matter of right; eligibility is fixed by rule; the bureaucrats handling their cases are meant to follow established procedures that are written at a safe distance from the counter. Unlike a private company, the bureaucracy cannot make up the rule as it goes. The word *compulsory* also highlights why bureaucratic arbitrariness is worse than that in a private exchange. A customer can take their business elsewhere, while a citizen cannot take their case to a competing state.

The word *nondiscretionary* means that in theory, what the citizen receives follows from legal duty and written rules, not from the goodwill of the person behind the desk. In practice, however, as the word *largely* highlights and as Lipsky (2010) has argued, discretion at the street level is widespread. Few laws are so determinate that they can be applied without any room for interpretation or judgment. This is for good reason. Citizens would want the law and its implementation made in a way that allows for customization to their particular case.

However, discretion in practice means that citizens subject to bureaucratic decisions can be at the mercy of some bureaucrat and their judgement. The citizen has no vote in this room for interpretation. The interpretation depends on another person who might as well have been somebody else. If the decision hinges on who decides rather than what is decided on, the citizen risks being "subject just to the *arbitrium*, the decision or judgement, of the agent", somewhat chosen "at their pleasure" (Pettit, 1997, p. 55). For a case, if another official would have made a different decision, the outcome depended on the allocation of the official, as the case might as well have received the opposite outcome. As the assignment of the official is impersonal and arbitrary by design, the case then also becomes, to some degree, arbitrary. Even if the rules left an official no choice in how to treat anyone inside a category, someone must still decide who is inside the category, and these decisions "are rarely clearcut": in making them, "administrators do make clients subject to their arbitrary will, to some greater or lesser extent" (Goodin, 1988, p. 322).

Whether the extent is greater or lesser is an empirical question, and it is one this dissertation aims to answer.

1.1 The bureaucrat and the citizen

The dissertation therefore examines the relation between the citizen and the state, and not other bureaucratic matters such as building a bridge or managing finances transparently. Throughout, I use words like official, bureaucrat, frontline worker and more, which all point to the person that in the end decides a citizen's case. In Lipsky's 2010, p. 3 words, I focus on street-level bureaucrats: social workers, judges, police officers and the like.

The study of the use of discretion, how much individual bureaucrats matter, has normally been examined through questions of how much bureaucrats are alike in reacting to different

case characteristics or common institutional pressures. This dissertation instead asks how much officials differ, facing similar cases. This can be described through a simple decomposition of error. Either errors are systematically in one direction (bias), or they deviate (variance). Whereas most prior studies look at which factors create biases on average, this dissertation looks at the variance and the more arbitrary deviations in bureaucratic decisions.

While this type of error in public decisions might be less studied, it is not new to the literature. The most pronounced scholar in the study of bureaucracy, Max Weber, compared errors from bias and errors from variance, when discussing what corrupts bureaucracies. Weber even treated variance as the more serious case. Whereas constant factors biasing decisions were in his mind at least "calculable in practice", he worried more about highly variable outcomes that are settled "from case to case with every individual official" (Weber, 1978, pp. 239–40).

Variance is therefore especially likely to produce hidden everyday, arbitrary injustices. Each decision can look trivial, correct and perfectly ordinary even though the citizen's outcome could easily have been different simply through allocation of the assigned bureaucrat. In the eye of the citizen, it might not matter whether their decision was wrongly handled due to biasing factors or due to simple bad luck in the lottery of bureaucrat assignment. What matters is that their case was wrongly handled.

There are thus two kinds of people involved in any bureaucratic encounter and as such two possible sources of error: (1) the official who receives the case and makes the decision, and (2) the citizen who must work the process to receive the treatment they are entitled to.

The official's discretion and the citizen's take-up are usually studied by separate literatures, but they function as two halves of the same question: does that which a citizen receives follow from their case and the law, or from the people involved? On the official's side, the right delivery of services and benefits can fail because the person behind the desk fills in what the rules leave open. On the citizen's side, it can fail because the person in front of the desk never applies, gives up during the process, or does not appeal a wrong decision. The two failures are also connected, because the citizen's decision to claim their entitlements depends on what kind of process awaits them. A benefit that hangs on an official's judgment asks the applicant to expose themselves to scrutiny with an uncertain result, and some will decide it is not worth it (Ko and Moffitt, 2024, pp. 26–27). Thus, the dissertation examines the phenomenon that citizens sometimes do not get the benefits and services they are entitled to. More broadly, it looks at how the delivery of benefits and services, sanctions and permissions can go wrong, by asking the wider question:

How often, and why, do like cases receive unlike outcomes?

The dissertation's answer to this question is made through four papers in the chapters 6-9. They are summarised in Chapter 5 on key findings. The papers each answer their own research question and can be read separately from each other. Table 1 and Figure 1 provide overviews of these.

The main paper of the dissertation is *Paper 1* and it provides measures of how much bureaucratic decisions depend on the official assigned to the case. Across 95 settings, I find that, at the median setting and under very conservative assumptions, at least 13% of cases were dependent on the bureaucrat deciding the case. Furthermore, reusing 13 conjoint experiments, I

find that even when officials judged the exact same case and were forced to choose, they disagree 40% of the time. Broadly, the findings show substantial arbitrariness in bureaucratic decisions from the allocation of the official.

Then, *Paper 2* asks whether algorithms can help reduce arbitrariness. It does so by comparing two different ways of using algorithms and AI in bureaucratic judgement and discussing their advantages in possibly reducing arbitrariness. *Paper 3* asks whether the length, complexity and specificity of the law help to guide discretion in a way so that there are fewer administrative errors, and measures this using fine-grained reviewed decisions that are mapped to specific legal provisions in a panel setup.

To look at the citizen-side of the bureaucratic encounter, in *Paper 4*, I ask why eligible citizens do not always receive their entitled benefits in a meta-analysis of 66 field experimental studies and show that missing take-up is widespread across the world and that citizens lack information and assistance in applying for benefits, as interventions strengthening those broadly have positive effects on take-up.

Table 1: The four dissertation papers

Paper number	Full title	Empirics
Paper 1	Arbitrary Bureaucratic Decisions	**Quasi-experimental and experimental variation in allocation of bureaucrats.** Re-analysis of 95 settings with as-if randomly allocated bureaucrats and 13 conjoint experiments with bureaucrats.
Paper 2	Noise in Administrative Decision-Making: Public Administration's Unseen Error and AI's Role in Enhancing Personnel Selection Consistency[a] *Re-submitted to Public Administration Review*	**Repeated ratings and human–AI ensembling.** Identical candidate profiles rated before and after common advice, paired with counterfactual human–AI ensembles.
Paper 3	Longer Laws, Better Service? Legislative Complexity and Bureaucratic Quality[b]	**Within-provision panel variation.** 112,663 reviewed complaints linked to changes in legal provisions
Paper 4	Increasing Take-up of Social Benefits: A Meta-analysis of Field Experiments *Published in Journal of Policy Analysis and Management*	**Meta-analysis of field experiments.** 233 treatment effects from 66 experiments across benefits and countries.

[a] With Mogens Jin Pedersen and Florian Keppeler. [b] With Jonas Krogh Madsen.

1.2 The concept of arbitrariness

The main part of the dissertation is centered around a concept of arbitrariness. A bureaucratic decision process is arbitrary to the extent that, holding fixed the relevant facts of the case, the outcome depends on which official receives the case. This counterfactual I call arbitrariness

throughout this dissertation. A bureaucratic outcome is thus arbitrary if it could have been different solely because another eligible official received the same case; in a binary setting, the same case can be granted at one desk and denied at another. A system or a setting can then be arbitrary to different degrees, since in some settings, all or most cases are dependent on the bureaucrat, and in others there are few or none. This essentially follows Mashaw (1983, p. 73), as he describes the commonest definition of an arbitrary system as "one in which the person who decides a particular controversy is more important than the merits of the case". The descriptive estimand of arbitrariness is thus whether or not, and to which degree, cases are decision-maker dependent.

In a sense this is similar to the concept of "noise" in Kahneman et al. (2021), which inspires the dissertation and which I actively use in *Paper 2*. I do not use that concept because, as the authors state themselves, it is a compound of different things; for example, what many would call "variance", they call "noise". As the papers in the dissertation have more precise estimands, the concept would conflate things that I actively try to separate. Precisely because the dissertation is mostly about the variance between bureaucrats, their concept of noise is too broad.

I could also have named it after Ramji-Nogales et al.'s *Refugee Roulette* and called it "bureaucrat roulette", "luck of the draw" or any such more cutesy phrasing. However, calling it arbitrariness and stating that cases can be arbitrary centers the focus on the citizen and their rights. Pettit (1997) uses the same word when he describes a welfare claimant as in the same situation as a wife exposed to violence, an employee unable to complain, and a debtor dependent on a lender's grace. Clearly the cases differ in severity, but they share the same structure: one person's position depends on another's capacity to interfere without them having an effective way to challenge that power. His administrative example is the claimant who is "vulnerable to the caprice of a counter clerk for whether or not their children will receive meal vouchers" (Pettit, 1997, p. 5). This is a fitting example of what is empirically at stake in this dissertation.

Fischman (2014) uses the word "inconsistency" which frames it less as a problem for the citizen and more a problem for the bureaucracy. So, while the word arbitrariness does not perfectly align with its use in Pettit (1997), it reflects the injustice for citizens in bureaucratic life.

Furthermore, I use arbitrariness instead of a word like "random", because it highlights the systematic part of what I am trying to argue; it is not that the outcomes are completely unpredictable, it is that the arbitrary allocation of bureaucrats matters, not just the merits of the case. I also do not look at the bureaucrats' or citizens' subjective experiences and feelings. Whether they themselves regard the bureaucracy as arbitrary and "kafkaesque" is a question for another time.

The word arbitrariness highlights a clear departure from the principles on which modern bureaucracies are built; namely common procedures and impersonal offices to make governments' use of power and distribution of benefits non-arbitrary, just and fair. Weber describes the bureaucratic ideal as formal equality and decision-making "without regard to personal considerations"; without those restraints, he writes, "the door would be open to arbitrariness" (Weber, 1978, p. 225).

1.3 Structure of the dissertation

In the next two chapters, Chapter 2 and Chapter 3, I will expand on the bureaucrat side of the main question as that is what three of the four papers are about. They provide independent contributions and function as a frame to understand the main question, its relevance, and academic history. In Chapter 2, I try to trace the research question of how much the individual bureaucrat matters in the study of discretion in political science and public administration, where the question of discretion was first posed. I try to summarise what part of the literature is problematizing it and how the question is important to the field. In Chapter 3, I expand on what I actually call arbitrariness and try to answer and make clear some common potential objections. I look at what the concept of arbitrariness means when compared to other problems in decision-making such as bias, discrimination, thresholds and aggregation of decisions in panels.

In Chapter 4, I write a synthesis of the papers by discussing potential strategies for heightening the congruence from Fuller (1964), i.e. minimizing the distance between the stated laws, implemented policies and what the citizens actually receive; these are guided by the three last papers.

In Chapter 5 I summarise the key findings of the papers and suggest implications for bureaucratic practice and future research.

Figure 1 shows how the four papers address the same problem.

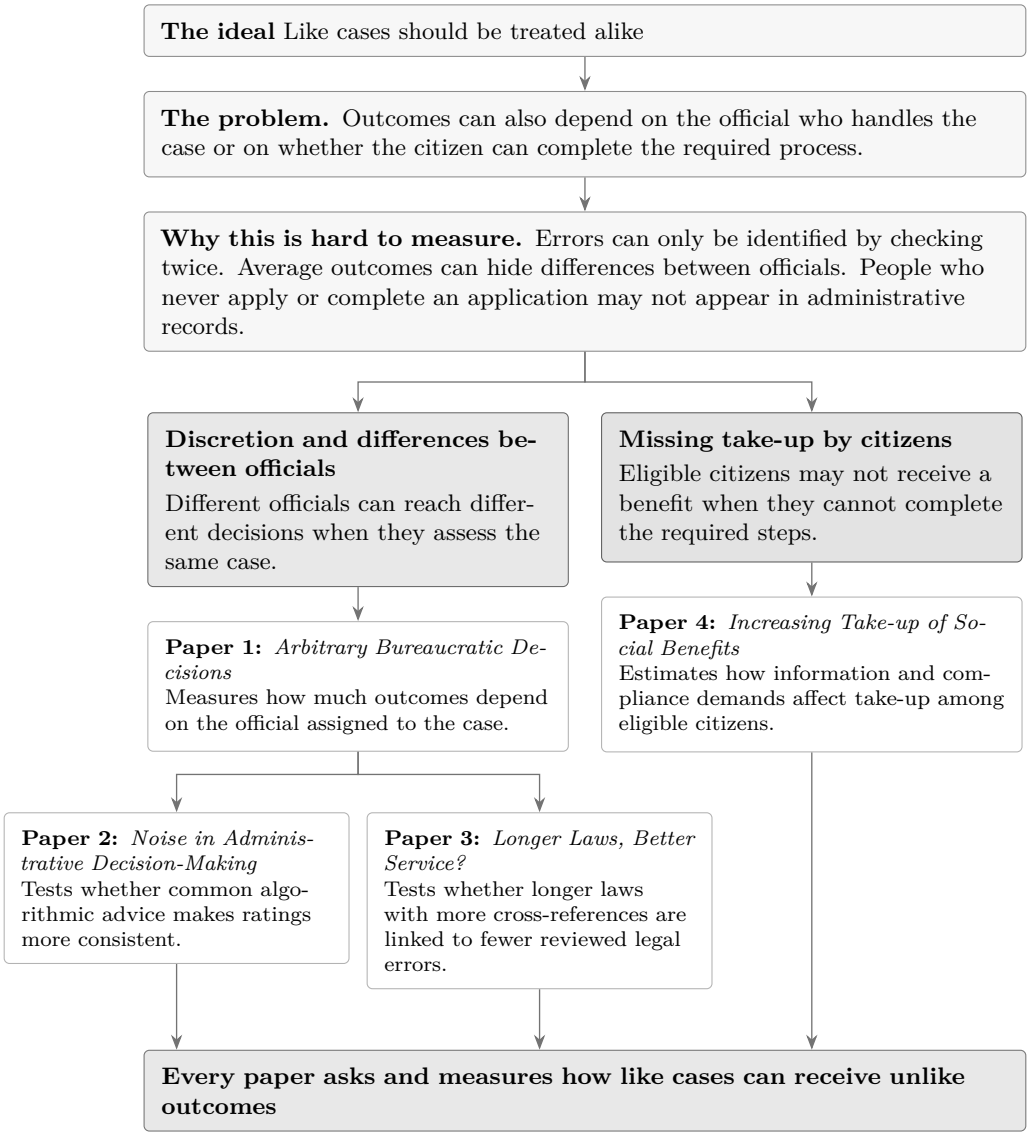


Figure 1: How the four papers address the dissertation's common problem.







Chapter 2

Discretion and arbitrariness

In this chapter, I give an outline of how the question of discretion has been approached in the study of bureaucracy. I argue that the literature has asked whether officials on average follow the standard set above them, far more often than it has asked whether officials at the same level decide the same case alike. The literature has thus focused more on the average than the spread; more on bias than on variance.

For the citizen, the problem of discretion has two dimensions, as shown in Figure 2. One is a vertical dimension asking whether the decisions of the delegated agents drift away from the top layer of the delegation chain, where the citizen has some say. This question is about whether each link in the delegation chain is consistent with the one above it. The other is the horizontal dimension asking instead whether each link is consistent at its own level. The first is about whether or not the public policies that street-level bureaucrats carry out are biased; the second is about whether they vary. This chapter attempts to show how the study of discretion came to focus on the vertical and not the horizontal question. The latter is the question around which the dissertation is centred.

I argue that the literature on discretion has moved away from actually investigating the fundamental problem of discretion as first formulated and towards describing the conditions under which frontline officials work and recasting discretion as a question of personal morality: what it means to exercise good judgment. Those questions are important and highly relevant, but they do not fully answer how much public outcomes depend on the bureaucrat and their discretion. Like the critiques by Gigerenzer and Brighton (2009) and Kahneman et al. (2021) on decision theory in general, I argue that the field has shown a "bias-bias" and thus moved away from looking at errors in decision-making and towards focusing on only one bias component driving them.

To give a roadmap, I argue that originally, the problem of discretion was that a bureaucrat's judgment stands between the citizen and the right to which they are entitled. The official can lean away from the standard (the vertical comparison, bias), or officials can differ from each other on the same case (the horizontal comparison, variance). The two forms both highlight the dependence on the bureaucrat. This was the problem that both Herman Finer and Carl Joachim Friedrich debated how to remedy. However, even though most authors writing on discretion mention arbitrariness and the horizontal problem as a vice, they do not focus on it. And the field of public administration almost never tries to measure it.

The aim here is not to give a comprehensive history of the study of discretion. There is a great tradition in the academic field of public administration of proudly reflecting on the field, so such a lengthy read can probably be found better executed elsewhere. However, I want to highlight some of the critical junctures in the literature that have guided the typical research questions towards questions of the structural conditions of street-level bureaucrats and how

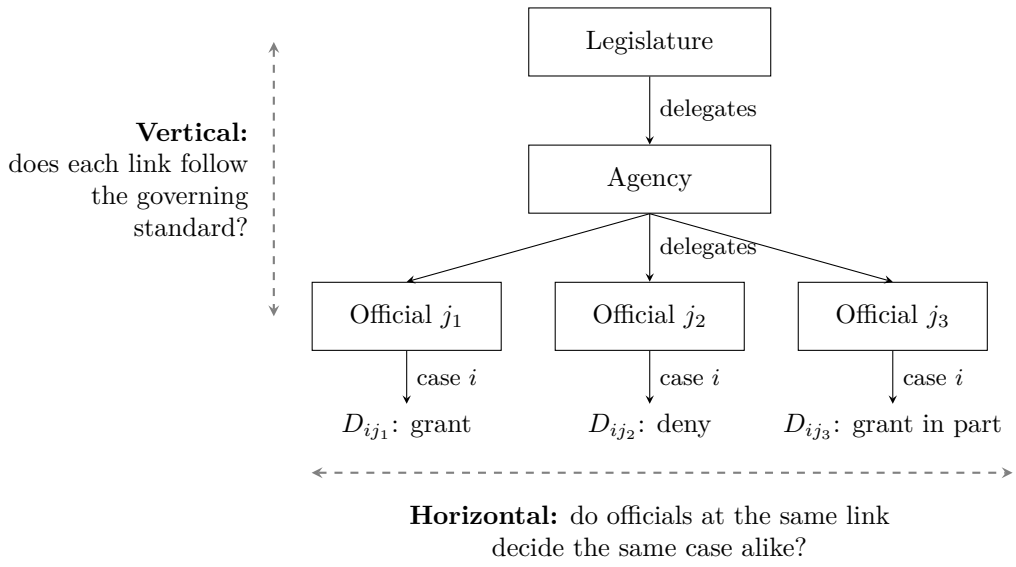


Figure 2: Two comparisons in the study of discretion.

they share moral dispositions, and away from questions of how faithful implementation actually is overall and how widely officials differ. So, in this section, I try to trace why the study of discretion has focused so little on horizontal coordination and consistency.

2.1 The original problem of discretion and the Finer-Friedrich debate

The field of public administration was partly founded on an idea about the separation between deciding and executing (Wilson, 1887; Goodnow, 1900), which later became known as the politics-administration dichotomy (Svara, 1998). Central to the idea is that the politicians give the administration discretion, but in turn expect their wishes to be carried out faithfully. The dichotomy was originally meant as a normative position, rather than a descriptive depiction of how the politicians and bureaucrats actually engage with each other, and in the normative sense the dichotomy has also survived the longest (Montjoy and Watson, 1995). The original problem can thus be understood in a simple chain of delegation like a principal-agent model, where the principals further up the chain need to ensure that the implementors as agents use their discretion in a way that aligns with the principals.

This line of thinking has a long tradition in the study of bureaucracy. It was the focus of implementation research primarily in the 1970s and 1980s, especially in the "top-down" accounts (see Matland, 1995) looking at how each link in the chain multiplied delay and failure (Pressman and Wildavsky, 1973). The question of discretion in general is thus bigger than that of street-level discretion—where the citizen meets the state. However, the normative discussion is similar.

In a Weberian understanding, the problem of discretion is immanent in bureaucracies. Hierarchy is preferable, and the whole reason for this is that the principals need the specialised knowledge the agents have, but that this same expertise makes it difficult for principals to specify and supervise the agents' tasks. The principals thus face a trade-off between controlling the

agents and trusting their expertise, or, in the words of Weber (1978, p. 994), the principals have to "exploit the special knowledge of experts without having to abdicate in their favor".

This was essentially the center of the famous debate on administrative responsibility between Carl Joachim Friedrich and Herman Finer in the 1930s and 1940s.

Finer thought that to have a democratic state is to ensure accountability to the elected politicians. As he wrote, representatives should determine official conduct "to the most minute degree that is technically feasible" (Finer, 1941, p. 336).

Friedrich went against the premise that administrators could simply execute one single settled public will. He argued that the sharp separation between politics and administration rested on a speculative idea that the state possessed a single will (Friedrich, 1940, p. 6). The "pious myth-makers" were, to him, those who took it for granted that formal answerability to parliament by itself "insures responsible conduct of public affairs by officials, high and low" (Friedrich, 1940, p. 3), and believers in subordination of that kind were "old-timers" (Friedrich, 1940, p. 19). In practice, administrators discover problems, choose means, and revise policy while carrying it out. Public policy is therefore "being formed as it is being executed" (Friedrich, 1940, p. 6). Friedrich therefore concluded that agreement between political principals and administrators would always be partial, and that democratic responsibility would remain fragmentary because the people spoke with an "indistinct voice" (Friedrich, 1940, p. 24). His claim was not that public preferences did not exist, but that they did not amount to one detailed set of instructions for administrators to follow.

How would policy then not drift towards the personal whims of the bureaucrats? How can bureaucrats be responsible? Friedrich argued that administrative responsibility had two conditions. Officials should answer both to the available technical knowledge and to popular sentiment (Friedrich, 1940, pp. 12–14). Technical knowledge was to be maintained through criticism by professional peers, while popular sentiment could be learned through the public's response to administrative policy via the parliament, the press, or contact with citizens. Having technical knowledge was not sufficient; popular sentiment was also necessary. Agreement among experts could not authorise a policy that the public opposed, but political direction could not turn a technically unsound policy into a responsible one (Friedrich, 1940, pp. 14–19). Friedrich borrowed John M. Gaus's phrase "inner check" for the civil servant's internalised responsibility to professional standards (Friedrich, 1940, p. 13). Friedrich thus placed part of the burden of securing responsibility on the individual bureaucrats themselves.

Finer rejected the claim that professional self-control could carry this burden because he thought it weakened the distinction between responsible judgment and an official's own sense of responsibility. Finer (1941, p. 335) thought this point was "extremely elementary" and mocked Friedrich's argument for placing responsibility in personal morals instead of political institutions, writing that Friedrich "pooh-poohs the efficacy of and need for political responsibility. He gives the impression of stepping over the dead body of political responsibility to grasp the promissory incandescence of the moral variety" (Finer, 1941, p. 339).

For Finer, responsibility required an external authority that could correct, punish, or remove the official (Finer, 1941, pp. 335–336). The public was after all still the master even when its wishes were unwise. He argued, that what mattered was not what officials believed the public

needed, but what the public and its elected representatives authorised them to do. Technical knowledge should inform political choice without displacing it. As Finer put it, the expert should be "on tap and not on top" (Finer, 1941, p. 348). He nevertheless accepted that instructions could not cover every case. Administrators needed some room for interpretation, and sometimes shaped policy themselves, but importantly their authority remained "conditioned" and "derived" within a hierarchy of narrowing discretion (Finer, 1941, pp. 340, 343).

For Finer, incomplete political direction would require more institutions and control, not only professional judgment. Friedrich accepted the need for political responsibility, but thought that external sanctions could not reach all administrative conduct. For Friedrich, the inner check was therefore part of the system of control; for Finer, it remained a useful disposition that could not establish answerability. So, in Friedrich's defense, his position was probably less permissive than Finer's critique suggests. Responsible policy had to meet two standards after all: technical knowledge *and* popular sentiment. Professional criticism and responsiveness to the public therefore had to supplement political control.

For the argument of this dissertation, the view on discretion could seem to mimic some of Finer's critique and scepticism. As he would argue, public outcomes should not depend on a bureaucrat's unsupported judgment, and appealing to professional responsibility does not by itself show that bureaucrats are sufficiently kept in check. Yet for the dissertation, the important point is that they shared the same question: how can we make sure the bureaucrat is responsible? How can we control the level of discretion of the delegated agents? Finer was interested in whether an official followed political direction; Friedrich was interested in whether a policy adhered to technical knowledge and popular sentiment. They disagreed about how much risk responsible administration could tolerate, but both assessed bureaucrats against some common standard. In that way, the original debate on discretion is a debate on two remedies to the problems of discretion: tighter, more determinate rules from above, or professional judgment from within. The debate is thus not about the average bias of the bureaucrats but the total amount of error—or the total distance from the two different standards over which they disagreed.

The horizontal comparison is therefore central to the debate: Rules are worth their cost only if bureaucrats left to themselves would differ, and professional judgment will result in even treatment only if professionals converge.

It is thus not clear whether the Finerian attempts at increasing alignment with the political principals—e.g. making the laws more specific as in *Paper 3*—are better at reducing arbitrariness than more Friedrichian policies, such as peer advice (or AI advice as in *Paper 2*), training, and stronger public norms. Resh (2017) writes that generally the positivist study of discretion has mostly followed Finer. This is apparent in public choice and principal-agent theory that treat large bureaucratic discretion as a loss that needs to be minimized (Calvert et al., 1989; Brehm and Gates, 1997; Niskanen, 1971). The studies that do follow Friedrich ask the vertical question, whether professional peer discipline can constrain bureaucrats whom political sanctions cannot reach (Miller and Whitford, 2016, pp. 141–154), or an "ethics of office" can (Møller et al., 2022). In the terms of figure 2, all of these theories are vertical: their agent is one actor, an agency or "the bureaucrat" in the singular, and the measures are at the level of the statute, the agency, or the average official.

The horizontal question is not asked as a problem of discretion even in the models that actually describe differences between types of bureaucrats. For example, Gailmard and Patty (2007) make discretion the currency a legislature pays with when it cannot pay in wages: "discretion is a substitute for money, but only for zealots" (Gailmard and Patty, 2007, p. 880), so the officials who invest in expertise and stay are the ones who care about bending policy their own way (Gailmard and Patty, 2007, p. 884). If discretion pays because it lets an official bend policy towards their own ideal, then, unless every zealot shares one ideal, two of them will bend the same case in different directions.

To examine the horizontal question and move beyond the theories that follow Finer or Friedrich (explicitly or not), we should look beyond theories of control and towards theories that operate more on one horizontal level. To look at the horizontal problems and the question of arbitrariness, like cases and unlike outcomes, we need to understand what happens at the link of the delegation chain and at the street-level bureaucracy.

2.2 How street-level bureaucrats actually are

Probably the most seminal book on street-level bureaucrats is the one named after the term, by the author who came up with it; in *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*, Michael Lipsky argues that most people do not really experience policies; they only experience concrete interactions with the government and the people representing it. Lipsky makes the key argument that before we discuss whether or not street-level discretion is good or bad, whether we should have more or less, we should accept that there will always be some discretion, and as such we should understand how policies are actually being carried out in order to understand what policy is and is not. However, whereas Lipsky was right in this observation, most of his book concerns the effects of structural features, the shared pressures of street-level work and as such, what street-level bureaucrats have in common. The book has thus helped centralize street-level bureaucracy as an important empirical object, but it does not systematically measure variation within it: citizens' treatment may also depend on which individual bureaucrat handles their case.

What Lipsky saw

Instead of looking at how policies are designed, we can simply look at how they are distributed, since that is where they have an effect anyway. Whereas prior implementation literature follows policy from its formulation and down the delegation chains of each administrative layer, Lipsky (2010) argues that what really makes a policy is what happens when people actually encounter it. Policy consists, in a sense, of the decisions that ultimately go through specific people in charge of distributing benefits, services, sanctions and permissions. Importantly, these people have a lot of discretion and many of the decisions are thus at their mercy; the decisions then "become, or add up to, agency policy" (Lipsky, 2010, p. 3).

Many of these institutions could be understood as "coping" organizations (Wilson, 1989);

both the outputs and outcomes are hard to measure and as such, we rely on the expertise of bureaucrats with the delegated authority to deliver the policy. A good example is how schools work. Neither teachers' core work nor how that contributes to student learning can be observed cleanly and clearly. For the observability of ouput, it is possible for a principal to observe a lesson, but the visit may change how the teacher behaves; for the observability of outcomes, we could use test scores, but those combine teaching results with the pupil's prior ability and the parents' willingness to help after school. Street-level bureaucrats' role is therefore hard to measure and as such hard to control and guide. We must rely on their expertise and trust that they use their discretion properly, at least to some degree.

In explaining street-level bureaucracy, Lipsky (2010) starts from the shared conditions of this type of official. For these people, resources are chronically inadequate relative to the work, because the demand rises to meet the supply, as most government benefits are publicly provided private goods. Expanding access to schooling, mental-health treatment, or legal assistance would typically just increase the demand for schooling, mental-health treatment, or legal assistance (Lipsky, 2010, pp. 33–38). These organizations also have goals that are unclear or conflict with one another. Schools, for example, both have to respond to each pupil's needs while teaching common skills and maintaining order (Lipsky, 2010, pp. 40–44).

Frontline officials must therefore decide with limited time and information. To do so, they narrow what they count as their job and simplify how they see clients (Lipsky, 2010, pp. 82–83). In a sense, this can be understood through Simon's (1997, pp. 118–119) description of administrative behaviour as bounded rationality: bureaucrats focus on a few aspects of the situation and consider only some possible actions and then choose a course that is "good enough" rather than identify the best among all alternatives. These responses turn into routines that can determine policy within the bounds set by their institution, sometimes with outcomes that depart from its stated aims (Lipsky, 2010, pp. 83–85).

Given these conditions, Lipsky (2010) devotes the later chapters to concrete tactics that street-level bureaucrats employ. They ration scarce services and impose the costs of access unevenly (Lipsky, 2010, pp. 87, 95). They cream-skim in that they prioritize clients who look likely to meet the agency's measures of success, use practices that keep clients under control, and process people through stereotypes (Lipsky, 2010, pp. 107–117). Lipsky does not treat these practices simply as failures of worker character. Instead, he describes them as structured coping mechanisms. This is why his stated object is the "common patterns of practice" produced by these shared work conditions and "the direction these patterns take" (Lipsky, 2010, p. 27).

What Lipsky did not see

Because so much of Lipsky's work focuses on the shared factors, Baviskar and Winter (2017) argues that it is not a theory of how the individuals in public administration actually matter. To really look at how the individuals matter, we need to look at the differences between street-level bureaucrats rather than their similarities.

It is not that differences among officials are impossible within Lipsky's theory. Shared workloads, ambiguous goals, and scarce resources create a common need to cope, but they do not

determine how each official will do so. The routines officials adopt can be affected by personal biases, beliefs, skills and so on. Different officials may also simplify the same task in different ways (Lipsky, 2010, pp. 84–85). Between-bureaucrat variation is therefore a natural inference from Lipsky's theory.

Lipsky's argument is that discretion can be valuable because of what it allows officials to accomplish. It can be thought of as having compensating benefits when it lets an official notice a relevant feature of a case, respond to a client's needs, or avoid the harm that would follow from applying a general rule mechanically. If unequal treatment produces none of these gains, it is hard to defend. Lipsky therefore argues that discretion should be restricted when it leads to "unfair and unequal treatment of clients" without any offsetting benefit, calling such treatment an "unredeemed source of unfairness" (Lipsky, 2010, pp. 196–197). His position is thus still that discretion and the potential for unequal treatment require justification.

Lipsky later explains what such a justification might look like. Public services must treat similarly situated people alike, but officials must also be able to respond to important differences in the cases. For example, a social worker may temporarily overlook missing documentation when rigid enforcement would deny an otherwise eligible person assistance. Lipsky argues that good rules should leave room for exceptions when they can be defended on the basis of the presenting facts. Exceptions based on social ties, ethnicity, or favoritism instead undermine trust because they cannot be justified in the same public terms (Lipsky, 2010, pp. 229–230). Therefore, his argument contains the basis for distinguishing between variation across cases and variation across officials. Different treatment can be justified when relevant differences between cases call for different responses. Lipsky would thus agree that like cases should have like outcomes.

It is, however, clear from the examples he uses that he is primarily interested in describing common tendencies. For example, he describes a study on tonsillectomy. A panel of physicians recommended tonsillectomies for 174 of 389 children. A second panel then examined the 215 children whom the first panel had passed over and recommended the operation for another 99. A third panel examined the remaining 116 children and recommended the operation for nearly half of them (Lipsky, 2010, p. 113). The later panels therefore recommended treatment for children whom earlier panels had examined and rejected. This example contains direct disagreement over the same cases, the same children; this could have been an alternative reading of the study. However, Lipsky uses the example to make a different point. He argues that service professionals will continue to find some clients who appear able to benefit from their intervention, even in populations other professionals have already screened. His inference concerns what professionals have in common, while the design also contains visible disagreement between them. That example thus contains two distinct, orthogonal questions: what shared conditions lead the physicians to act as they did, and how much they actually differed in their assessments of the cases.

We can thus easily extend and use the theories of Lipsky (2010) to ask whether coping behaviour actually leads bureaucrats to behave alike or differently.

This leads us back to the problem of accountability, responsibility and control of bureaucrats. Lipsky writes that an official is accountable when there is a high probability that the official will respond to legitimate authority, and only a pattern of behaviour can make conduct predictable

enough to judge (Lipsky, 2010, pp. 159–160). Holding street-level bureaucrats accountable therefore requires four things: it must rank what it wants its workers to do, measure their performance, compare workers with one another to establish a standard, and be able to enforce it. Importantly, notice the third condition; someone has to put comparable workers next to one another on a common standard. Normatively, Lipsky actually makes the horizontal comparison of Figure 2 a requirement for accountability. However, Lipsky argues that these conditions are "virtually impossible" to meet at the frontline, because the work also demands openness to each case, so officials answer to clients as well as to the agency (Lipsky, 2010, pp. 159–162).

Lipsky showed that street-level discretion is important, but that it can neither be removed nor fully controlled. This is in a sense more in line with Friedrich. If bureaucrats have this large amount of agency, the question in the literature naturally becomes a question of personal morality; if the system is an aggregation of the people involved, the quality of the policies then necessarily follows from what it means for the bureaucrat to exercise the judgment well.

2.3 Are there right answers?

Now we know that it is contested how to constrain discretion appropriately, and we know that it matters. How then should it be used? What are the virtues of good administrative decisions? And what are the vices if arbitrariness is not the primary one?

Instead of asking whether an administrator is responsible, like Finer or Friedrich, Simon (1997) argued that we should start at the decisions themselves rather than at thoughts on the role of the administrator. We instead should examine what a decision is made of. Every administrative decision, he argued, mixes two kinds of premises: (1) factual premises and (2) value premises. Factual premises say what the world is like, and they can be true or false; value premises cannot. As he writes: "If two persons give different answers to a factual problem, both cannot be right. Not so with ethical questions" (Simon, 1997, p. 62). Although he concedes that there are indeed some parts of administrative decisions that cannot be called correct and that somebody must make a value judgment, he is not Friedrichian. He argues that "self-responsibility of the administrator is no answer to the problem" and that value judgments are validated by democratic procedure, and "expertise of whatever kind is no qualification" for making them (Simon, 1997, p. 65). He does not view value judgments as something the individual bureaucrat should provide; these were meant to be provided outside the decision, by the legislature and transmitted downward through the structure.

Whether or not a bureaucratic decision can be called correct is of course central to the dissertation. If there is no right decision then there is no wrong either. This is also one of the most fundamental discussions in legal philosophy on the question of whether or not the law is always determinate or sometimes indeterminate. That the law is determinate means that even for hard cases, the facts of the matter will always require a unique decision, whereas if the law is indeterminate, it leaves more than one disposition admissible, so that the same case could be decided either way and neither decision would be overturned upon review. Theoretically, and excuse the very short summary of one of the most important discussions in legal philosophy, legal

positivists, most notably Hart (1994), argue that general rules have a core of settled meaning and then a "penumbra" of doubt, and that this openness comes from a limit "inherent in the nature of language", so most cases are plain and a minority, hard cases, are then left for officials to use their discretion to settle. By contrast, Dworkin (1977) argues that even hard cases (in principle) have a right answer, reached from the principles the legal materials already contain, so that officials must judge without ever being free to choose. If two officials decide the same case differently, according to Dworkin, one of them is wrong; according to Hart, neither is.

In empirical terms, Fischman (2014) argues we can thus not easily separate indeterminacy from error from officials' decision rates (such as the ones used in *Paper 1*). To determine whether there has been an error thus requires checking twice and treating the higher institution as right; this is what *Paper 3* utilizes, as do other papers in the field measuring "errors" (Greer and Bullock, 2018; Young et al., 2024; Widlak and Peeters, 2020).

When is this judgment acceptable? Simon (1997) wrote little on it, although he noted that in case of disagreement between the administrator and the legislator, the former must have "complete answerability" (66). In a sense, he is thus more in line with Finer; however, he is aware that administrators have personal values that they would like to see implemented. As such, he essentially argues that it is in the nature of administrative decisions that they can be arbitrary, although he did not condone it. This is in line with procedural arguments that we can accept a decision based not on the outcome but on the process.

For example, Davis (1969) wrote that we should design discretion to be confined, structured, and checked. Another proponent of such a view is Mashaw (1983). Similar to the empirical argument in *Paper 1*, Mashaw (1983) noted that there were large dispersions in an American disability agency. From this, he made a larger theory of administrative justice, which in his words is "those qualities of a decision process that provide arguments for the acceptability of its decisions" (Mashaw, 1983, pp. 24–25). That seems a bit tautological and vague, but the important part here is that he writes in the plural. Specifically, he argues that there are three different benchmarks used to judge a decision. (1) "Bureaucratic rationality" judges decisions by their fidelity to the stated policy and their cost. (2) "Professional treatment" judges them by whether expert judgment serves the individual client. (3) "Moral judgment" judges them by whether a neutral process gives claimants an equal chance to secure their entitlements (Mashaw, 1983, pp. 25–31). The three benchmarks are "not mutually exclusive" but "highly competitive" (Mashaw, 1983, p. 23).

For bureaucratic rationality, the central vice is arbitrariness where an outcome depends on who decided (Mashaw, 1983, p. 73). For professional treatment, two officials can reach different views about what will help a client without either view being wrong, provided each applies professional standards to those circumstances of the case and seeks to serve the client (Mashaw, 1983, pp. 27–29). For moral judgment, he argues that sometimes legal standards cannot settle hard cases on their own, a position more in line with Hart (1994) and against Dworkin (1977). The decision-maker must then interpret what the open-ended standard requires, so the decision both resolves the claim and defines what the claimant is entitled to (Mashaw, 1983, pp. 29–31).

I will return to the normative underpinnings of the argument of the main question guiding the dissertation, especially in *Paper 1* and *Paper 2*. However, no matter which position one

takes here, there is no position in which arbitrariness is completely tolerated. It is always seen as a vice, albeit potentially with a tradeoff. Whether or not the law is determinate and we thus in theory can point to an error does not matter for the dissertation. The question is not whether a decision would survive a review, but whether it would have been the same in somebody else's hands.

We have now established that street-level bureaucrats use discretion widely, and it is in value judgments, which are an inevitable part of street-level work, that they are most likely to differ. This is the dissertation's main focus. As such, if any authors studying discretion could have discussed the horizontal question, it should be the ones looking at how personal, individual decisions are viewed in the literature of street-level bureaucracy. How do the street-level bureaucrats themselves describe their use of discretion?

2.4 The moral street-level bureaucrat

Two ethnographic books have shaped how scholars understand discretion and value judgment at the frontline: Steven Maynard-Moody and Michael Musheno's *Cops, Teachers, Counselors: Stories from the Front Lines of Public Service* and Bernardo Zacka's *When the State Meets the Street: Public Service and Moral Agency*. Both start off by rejecting theories positing that we can view street-level bureaucrats as only the last part of the delegation chain following orders, both the normative and the descriptive ones. Zacka writes that he rejects the "compliance model" of bureaucratic responsibility (Zacka, 2017, p. 36), a term from Brodkin (2000). He explicitly presents his argument as a return to the debate between Finer and Friedrich, siding with Friedrich (Zacka, 2017, pp. 33–35). Maynard-Moody and Musheno call it the "state-agent narrative." Neither book rejects that law, hierarchy, or accountability matter. Their claim is that this does not cover the full picture because frontline workers also understand themselves as "citizen agents" who make judgments about the people before them (Maynard-Moody and Musheno, 2003, pp. 9–13), or draw on moral dispositions in order to overcome the conditions of street-level work (Zacka, 2017).

For Maynard-Moody and Musheno (2003), I think the key to their argument is their notion of using discretion as making "moral judgments"; following McCollough (1991), these are understood as "not simply 'emotive' expressions of how people feel, but cognitive claims about how things are, that is, moral knowledge", and they are in this respect "similar to scientific judgments" (Maynard-Moody and Musheno, 2003, p. 155). This obviously goes against the ideas about value judgments from Simon (1997), as they here group the two categories he purposefully separates, factual and value judgments.

They describe how the frontline workers judge the identity and moral character of the person in front of them, and then "apply, bend, or ignore rules and procedures to support the moral reasoning" (Maynard-Moody and Musheno, 2003, p. 155). For them, fairness does not mean treating everyone the same, but instead responding to a person's worth (Maynard-Moody and Musheno, 2003, pp. 93–94). It is not that these judgments are inconsequential. They decide who gets years of schooling and who gets a referral to a minimum-wage job, who is arrested

and whose crime is ignored (Maynard-Moody and Musheno, 2003, p. 106). And the book is actually aware of this dissertation's object, which they state as a plain fact: "In story after story, street-level decisions lack the consistency and coherence to be called policy. A street-level worker may handle similar cases in different ways, and workers from the same jurisdiction may handle cases differently" (Maynard-Moody and Musheno, 2003, pp. 11–12).

Interestingly, the authors do not condemn this. For much of the book, they seem to admire it. A teacher who works the consent paperwork so that her student keeps services a budget cut denies to similar students is "savvy" (Maynard-Moody and Musheno, 2003, pp. 110–111). Yet they do not fully celebrate it blindly either. As they write: "there are no unambiguous criteria to discern citizen-client worthiness" (Maynard-Moody and Musheno, 2003, p. 119); once a client is labelled a bad guy, "every infraction of the rules, every character flaw, is used to limit service" (Maynard-Moody and Musheno, 2003, p. 143); too much of the judgment, they write, runs at the unreflective, expressive level (Maynard-Moody and Musheno, 2003, p. 160); and they write that street-level bureaucrats' power advantage over citizens should not be exploited (Maynard-Moody and Musheno, 2003, p. 159). To remedy this, they suggest peer review between bureaucrats.

Reading the book descriptively, it would seem natural to extend it and examine differences in how frontline workers draw on ideas of deservingness. Reading the book normatively, they condone differences driven by "worth" in the best cases and note that they are wrong in the worst, but say little about how to tell the two apart.

Zacka (2017), unlike Maynard-Moody and Musheno (2003), uses the factual-value distinction from Simon (1997). He argues that frontline work is moral work in all of the challenges that make street-level work not fit the "compliance model". As such, for each of his challenges, he writes that officials must give practical content to ambiguous goals, choose between conflicting goals and values, decide how limited resources should be distributed, draw lines across fuzzy boundaries, adopt rules for which errors to tolerate under uncertainty, judge soft evidence that cannot be reduced to measurable facts, respond to situations that could not be predicted in advance, choose means that may alter the ends being pursued, and rely on their own sense of responsibility where information asymmetry and moral hazard make complete supervision impossible (Zacka, 2017, pp. 48–63).

Zacka defends frontline discretion because officials must balance efficiency, fairness, responsiveness, and respect. His defence is conditional on keeping all these demands in line (Zacka, 2017, pp. 99–101). He actually explicitly mentions arbitrariness as a risk to be controlled, writing that discretion "raises the specter of arbitrary treatment" (Zacka, 2017, p. 14). Zacka therefore treats protection against arbitrariness as a central purpose of rules. Following Goodin (1988), in a footnote, he explains that it is desired that officials follow established rules and deliver services to which citizens are entitled as a matter of right, and that those who make the rules are purposely at a distance from service provision precisely to insulate recipients from potentially arbitrary frontline decisions (Zacka, 2017, 101 n. 51). However, rules can produce their own arbitrariness. Zacka therefore describes it as a trade-off: some inconsistency among frontline officials may be tolerated when the alternative is a rigid and equally arbitrary cutoff point (Zacka, 2017, p. 56).

When Zacka describes what street-level bureaucrats do, he therefore describes pathologies.

He does not mean that frontline moral judgment is itself pathological. His three ideal-typical pathologies, indifference, caregiving, and enforcement, become pathologies when a worker treats one valid part of the role as sufficient and loses sight of its other demands. The indifferent follows routines without exercising the judgment the role requires; the caregiver sees the needs of one client but loses sight of fairness across cases; and the enforcer sees abuse and the protection of public resources but becomes blind to need and respect. Each may remain well intentioned and act within the bounds of the role while viewing the case through a narrowed moral field (Zacka, 2017, pp. 96–110). Zacka thus reinterprets Maynard-Moody and Musheno's interviews: the workers go wrong not because they make moral judgments, but because one moral disposition screens out competing reasons (Zacka, 2017).

Looking at the individual decisions of the bureaucrats and judging whether or not they follow the appropriate moral dispositions and so on is a valuable endeavour. However, arbitrariness is a problem of the system and not the individual bureaucrat. Personal morality cannot answer a system-level question. An official may reason carefully and still reach a different outcome from an equally careful colleague facing the same case. Focusing on moral appraisal of the individuals is to ask whether the official's reasons are defensible; the problem of arbitrariness asks whether the outcome depends on which official receives the case within the system. Finer called his own distinction between a sense of responsibility and effective answerability "extremely elementary," while observing that "a writer of eminence" had discounted its significance (Finer, 1941, p. 335). The corresponding point here is equally elementary: arbitrariness cannot be assessed by looking at averages, at how bureaucrats are alike, or at one bureaucrat at a time.

2.5 Does the discretion literature measure arbitrariness?

I have now shown that street-level bureaucrats matter and that they make different value judgments in their job. Arbitrariness is a vice in the original formulation of the problem of discretion, whether or not it is called the compliance model or the state-agent narrative, but it is similarly a vice in theories of bureaucrats' ideal moral dispositions. Arbitrariness is therefore only sometimes tolerated when it is traded off against other values.

The next logical question is therefore to ask whether the empirical literature on discretion has actually measured arbitrariness. When the public-administration studies examine individual officials who decide or handle cases, how often do they report how widely those officials differ?

To answer that, for the purposes of this chapter, I coded the field's two core journals using large language models. I took every article the *Journal of Public Administration Research and Theory* and *Public Administration Review* published between 2010 and 2025. I found 304 that a keyword search marked as work on street-level bureaucracy or on compliance, and found 60 papers that were actually studying individual frontline officials deciding or handling cases. I have focused on these two journals because they are widely regarded as the top outlets in the subfield. This census is not comprehensive, but here it functions as a way to structure different themes. The answer is that it almost never does.

For each of the 60 articles, I examined four questions. (1) Does the paper claim that officials

differ, and if it does, does it estimate how far? A claim and an estimate are not the same thing. (2) Were the officials it compares looking at comparable cases, or simply at different ones? (3) When the paper reports a difference, is it a difference between kinds of citizens or a difference between the officials themselves? The first is a statement about the average and the second about the spread. And (4) where a spread between officials is measured, does that number bear on what the paper goes on to recommend? All questions are meant to assess whether the field built on what individual bureaucrats do has ever checked how much the individual official turns out to matter.

Five of the 60 papers report a measurement of how far officials facing comparable cases differ. Two find that officials differ (Keiser, 2010; Schmidt et al., 2025). Two report the spread and never say whether it is large (Jensen and Pedersen, 2017; Halling and Petersen, 2024). The fifth finds officials alike in sharing a bias against the same clients (Shoub, 2022).

Most closely related to this dissertation, Keiser (2010) studies randomly assigned examiners in the American disability programme. She finds that examiners report allowance rates with a mean of 33% and a standard deviation of 17 points, with a total spread running from 2 per cent to 65 per cent (Keiser, 2010, p. 252). This is the cleanest measurement in the corpus and the only one of the five that is not a byproduct of a study of client-group bias. Its weakness is that the rates are self-reported rather than drawn from real administrative data and the actual files; otherwise, it mimics the numbers produced in the first part of *Paper 1*, looking at the difference between bureaucrats harvested from leniency IV studies.

Schmidt et al. (2025) give identical vignettes to frontline professionals and ask when they would bend a rule for a client. In the paper, they are primarily interested in how the frontline professionals are alike, but they look at random intercept slopes, and they find that about two thirds of the variance that remains once the case is accounted for is due to differences between professionals (Schmidt et al., 2025, p. 373). The authors call this variation substantial. But their question is which client gets the rule bent, not how much professionals differ, so the dispersion is not something they really are interested in.

Two more studies produce relevant numbers, but they never say whether they are large or small. Jensen and Pedersen (2017) look at a conjoint experiment of disciplinary cases with prison employees and fit random coefficients, so employees may differ not only in how strict they are overall but in which attributes they weigh. The standard deviations across employees run from 0.64 to 1.30 and are of the same order as the average effects they sit beside. This looks a bit like the conjoint part of *Paper 1*, although it is not interpreted as such.

Halling and Petersen (2024) give caseworkers identical vignettes, and under a table of descriptive statistics they note a between-caseworker standard deviation of 1.98 on a scale from zero to ten, beside a within-caseworker standard deviation of 2.27. The words variance, dispersion or the like do not appear anywhere in the article, and the discussion does not mention that caseworkers differed.

Shoub (2022) compares police officers' search threshold for black and for white drivers. Between 79 and 83 per cent of officers search black drivers on weaker grounds. Here, they look at some of the extreme officers as an outlier test, and they argue that even when removing the outlier officers, there is an effect on average.

Taken together the five say something about the field beyond their own findings. Public administration typically measures bias on purpose and variance by accident.

The other 55 papers ask different questions, and the questions fall into three groups. I will now shortly summarise those groups since it is key to understand what has been perceived to be important in studying discretion and street-level bureaucracy. The studies look at either (1) discrimination of citizens, (2) bureaucrat type, or (3) structural features of being a street-level bureaucrat. In my summary, I will mention some studies other than the 55 to give a better picture of the questions at hand.

Table 2: The census mapped onto the four coding questions

Question	Result	Studies
1. Is between-official spread measured?	Yes	5
	No	55
2. Are the five studies based on comparable cases?	Identical or standardised cases	3
	As-if random assignment	1
	Risk-adjusted cases	1
3. What is compared across the full census?	Individual officials	5
	Citizen groups	10
	Observable bureaucrat types	9
	Structural features of frontline work	36
4. Does the measured spread inform recommendations?	Yes	2
	No	3

Note: Questions 1 and 3 cover all 60 papers. Questions 2 and 4 are conditional on the five papers that measure between-official spread.

Discrimination of citizens

Many studies look at whether bureaucrats react alike in treating citizens of various races, genders, or ethnicities differently. These studies estimate whether citizen characteristics such as race, ethnicity or religion of the citizen shift the average probability of an error, punishment, access, or assistance (Compton et al., 2023; Olson, 2016; Andersen and Guul, 2019; Xu and Lee, 2025; St. John et al., 2026; Pfaff et al., 2021). We can think of this as a directional gap from an appropriate decision, a mean shift. Since they measure the average effects of whole institutions, this does not show that citizens' outcomes depend on which official they draw.

These studies are in a way similar to another group of studies that asks which client behaviours strengthen their chances of receiving further help (Jilke and Tummers, 2018; Guul et al., 2021; Mikkelsen et al., 2023; Lavee, 2021). These are typically survey experiments that hold the official constant, but vary the client traits, and interpret which features of the case move the average decision.

No administrative error is better measured than bureaucrats' bias against particular kinds of citizens. Outside the census, this is the same research question in audit and correspondence experiments that look at actual public encounters. The method is familiar from the canonical

labour-market study by Bertrand and Mullainathan (2004); in studies looking at public services, they typically randomize client cues in requests to election officials, housing agencies, benefit offices, or primary schools (White et al., 2015; Einstein and Glick, 2017; Hemker and Rink, 2017; Giulietti et al., 2019; Olsen et al., 2022). A systematic review identifies 90 public-administration-relevant articles using the audit design (Beasley and Lahey, 2024). Similar to the survey experiments, these studies are directed at the case rather than at the bureaucrat. They ask how citizens are treated on average. The variation between officials is what they average over.

Bureaucrat types

A second group asks a restricted version of the horizontal question from Figure 2 by comparing observable types of officials. These studies examine whether treatment differs with the race, ethnicity, gender, or experience of the bureaucrat (Headley and Wright, 2020; Pedersen et al., 2018; Oberfield and Incantalupo, 2021; Baker and Hassan, 2021; Zhang, 2025). They show that the citizen's outcome can sometimes depend on the specific kind of official who gets to decide the case. For example, Headley and Wright (2020) find that officer race conditions the use of force against black civilians, but that the racial disparity in arrests is equal across officer race.

These studies do show how some people depend on the type of bureaucrat, but they do not measure the full spread, since officials can differ within each type of bureaucrat and because equal average biases across types can cancel out and conceal substantial person-to-person variation.

A few studies come closer to measuring that spread. Harrits (2019) looks at whether each official applies a "class stereotype" in a vignette experiment and shows that this pattern varies with the social context of the workplace. Chang (2026) examines police officers through successive encounters and finds that their treatment of same-race drivers changes with what happened in the preceding encounter. Friedrich and Eckhard (2026) report how much variation in speech there is in conversations between citizens and bureaucrats, while Adams et al. (2021) report a wide distribution of body-camera activations across officers. Each study measures variation at the individual level, but they do not estimate the arbitrariness of the bureaucrat draw and do not make the argument.

Outside the two journals, many other studies have examined differences in bureaucrats' attitudes and background. To highlight some, studies of ethnic bias in adjudication show this, e.g. in Israeli small-claims courts, where cases were effectively randomly assigned, Shayo and Zussman (2011) find an in-group bias of Arab and Jewish judges, and similarly in Kenya Choi et al. (2022) find that judges are 3 to 5 percentage points more likely to grant an appeal when the citizen shares their ethnicity.

Footnotes to Lipsky

Famously, Alfred North Whitehead referred to the entire history of Western philosophy as "footnotes to Plato". A lot of public administration scholarship could in a similar way be described as footnotes to Lipsky, perhaps including this dissertation. Many studies closely follow Lipsky, in looking at the coping concept, the conditions of frontline work, and what

officials do when demand is larger than what they can supply (Tummers, 2017; Tummers and Rocco, 2015; Kaplaner and Steinebach, 2024; Løberg, 2023; Morrell and Currie, 2015; Ferraro et al., 2026). Related studies examine specific organisational tools or features to look at how they change bureaucrats' behaviour on average, such as monitoring, political control, formal limits on discretion, budget pressure, performance information, peer deliberation, administrative assistance, digital systems, and algorithmic advice (Amirkhanyan, 2010; Girth, 2014; Garrow and Grusky, 2013; Stensöta, 2012; Zhan et al., 2014; Kauppi and Raaij, 2015; Cárdenas and Ramírez de la Cruz, 2017; Boer et al., 2018; Møller, 2021; Visser and Hulst, 2024; Junge, 2022; Møller, 2025; Hjortskov et al., 2026; Selten et al., 2023; Headley et al., 2025; Su, 2021; Alcadipani et al., 2024).

A closely related set of studies looks at what officials do when the rules do not fit the case in front of them. Here on average, they diverge, deviate, repair, improvise, or respond (Gofen, 2014; Henderson, 2013; Masood and Nisar, 2022; Visser and Kruyen, 2021; Carlsson et al., 2026; Cecchini and Harrits, 2022; Pedersen and Pors, 2023; Maynard-Moody and Musheno, 2012; Cohen et al., 2023). It is largely qualitative and largely about what officials share i.e. how they are alike.

The administrative burden studies also belong under this heading. They look at what it costs citizens to apply for a service and stay enrolled (Kogan, 2017; Bell and Meyer, 2024; Döring et al., 2025; Porumbescu et al., 2026).

Other great studies outside public administration have also tried to measure causal effects in field settings that closely mimic Lipsky's theories. For example, Shumway and Wilson (2022) find that judges' caseloads increase as if randomly when peers retire and that the share of favorable decisions in disability insurance claims falls. Similarly, Kolstad (2026) studies a sudden increase in caseloads during COVID-19 among Danish unemployment caseworkers, who can adjust the number and timing of meetings, but finds no resulting increase in discrimination. Both studies also examine how frontline workers respond to the shared conditions of their work.

Thus, these are all within the strand of research that Baviskar and Winter (2017) criticised; a literature built to describe common pressures cannot report how far individual officials diverge, and is thus not about the individual.

What the grouping shows is not that the field lacks the data. Each of the three groups contains designs that could produce measures of arbitrariness from material the authors have already collected. A conjoint experiment with more than one decision per official can look at how bureaucrats are different. An audit study that records which official answered can compare distributions.

The vertical question is asked on nearly every page of the literature this chapter has covered. The horizontal question, whether it matters which official the citizen happens to get, is asked and answered five times in fifteen years of the field's two main journals, and where it is asked the answer is rarely the point of the paper.

2.6 Have other literatures answered the question?

There is, however, a large literature on public administration outside the field of public administration. And that literature is often quite good. I coauthored an article outside this dissertation, wherein we look at the field of economics and the public administration research it produces (Olsen et al., 2026). The argument can easily be extended to other fields,[1] but the paper shows that it is rarely sufficient to rely solely on the research in public administration journals when studying public administration questions.

The question of arbitrariness is fundamental to the problem of discretion as a democratic problem, but as I showed above, in the literature founded on that question, it is not so prevalent.

In economics, the question is taken up more often, strengthening the point that a lot of quality work on bureaucracy is published there. Best et al. (2023) look at what the Russian state pays for goods that are the same from one purchase to the next, so that any difference in price is a difference in how the buying was done. About a fifth of that difference can be traced back to the individual official who made the purchase, and about a quarter to the office they sit in, which means that which bureaucrat handles a purchase matters a lot. Bolhaar et al. (2020) study Dutch welfare applicants and record, caseworker by caseworker, which kind of help each applicant is granted. Some caseworkers almost never place a client straight into a job, while others do it for most of their clients, so the route an applicant takes through the system depends on the assigned official. Angelova et al. (2026) compare judges deciding who to release before trial against a consistent algorithm, which holds leniency fixed and leaves only the choice of whom to release. The algorithm is better than 90% of the judges when looking at which defendants go on to reoffend or miss court.

Prendergast (2007) makes a model of how biased a bureaucrat is towards a citizen, which essentially is the idea of leniency that *Paper 1* utilizes. In a formal model, he looks at what the salary level does to the distribution of bureaucrats' bias towards citizens.

However, in economics, more often the spread is the design rather than the finding (Chyn et al., 2025; Goldsmith-Pinkham et al., 2026). This is what I utilize in *Paper 1*.

Dispersion in adjudication

The literature that looks mostly at the same argument and also measures dispersion between officials is the literature on judges. Somehow, these studies are rarely referred to in public administration literature on discretion. Moving from caseworkers to judges may look like a change of subject, but Lipsky's own definition includes judges (Lipsky, 2010, p. 3). He argues that these otherwise separate occupations share "analytically similar work conditions" (Lipsky, 2010, pp. 3–4). Judges and caseworkers are not institutionally equivalent, but both translate a general public rule or law into a decision for a particular person.

Judicial research has studied the horizontal problem for much longer than public-administration research. Early studies compared the sentencing tendencies of individual judges with randomly assigned cases, and later work explicitly treated sentencing disparity as a

[1] See paopa.org

problem of unequal treatment (Gaudet et al., 1933; Frankel, 1973). Also, Partridge and Eldridge (1974) gave the same pre-sentence reports to 50 federal judges and compared the sentences they imposed. Fischman (2014) uses this study to discuss the difference between inconsistency, indeterminacy and error in decisions, focusing also on grant rates as in *Paper 1*.

The best known more modern versions are in refugee adjudication. Most notable is the work of Ramji-Nogales et al. (2007), who compare grant rates across United States asylum adjudicators hearing similar caseloads and conclude that an applicant's chance of asylum depends as much on the draw of a judge as on the case itself. The name they gave it, refugee roulette, became the name for the problem in other studies too. E.g. Rehaag runs the same comparison on Canadian Federal Court judges reviewing refugee determinations three times across a decade (Rehaag, 2012; Rehaag, 2019; Rehaag, 2024). This is the same person-dependence that the public-administration census above searched for among frontline officials.

I would argue that this is strange, since administrative discretion often is seen as less desirable than judicial discretion. Administrative discretion, as we have shown, was originally presented as something that must be explained or controlled because bureaucrats are expected to implement policy faithfully. Judicial discretion has greater institutional legitimacy because judges are expected to exercise independent judgment without referring to a manager.

2.7 Conclusion

For the original problem of discretion, both the vertical and the horizontal questions matter. Finer and Friedrich discussed how much trust we should put in the bureaucrats' moral judgment and as such whether or not the administration as a whole could be called responsible if the bureaucrats' personal whims were an effective part of delivering policy. This debate is thus not about average bias, but about the total amount of drift.

Descriptively, Lipsky showed that the decisions that street-level bureaucrats make are effectively what the policies are. However, he proposed to look at how bureaucrats are alike rather than how they are different. He was interested in the common structural conditions of how caseloads, ambiguous goals, and scarce resources produce a shared need to cope. Thereby, he was asking a vertical question of biases and not engaging with the question of variance.

To find authors looking at the horizontal questions, we thus needed to look at theories of decisions rather than the conditions of the bureaucrats. Only that can become a theory about the individual bureaucrat, why and how they can make different decisions. As such, the debate over whether right answers exist is foundational for whether or not we can think about differences in the decisions as merely moral differences that do not have an answer or actual errors. This is the starting point for Zacka (2017) and even he, who avidly defends street-level discretion as moral work, concedes that arbitrariness from differences between bureaucrats is only tolerable if it prevents other more rigid arbitrary rules.

As such, the chapter has shown that there are good normative and descriptive reasons for studying arbitrariness. It is important, undesirable, and presumably matters a lot; however, this type of research is absent in the field that builds on these reasons.











Chapter 3

What Is Arbitrariness?

In this chapter, I elaborate on what I mean by an outcome of a case being arbitrary. It is a big word, as it intuitively signals a strong departure from everything we associate with the benefits of bureaucracy, legality and the rule of law.

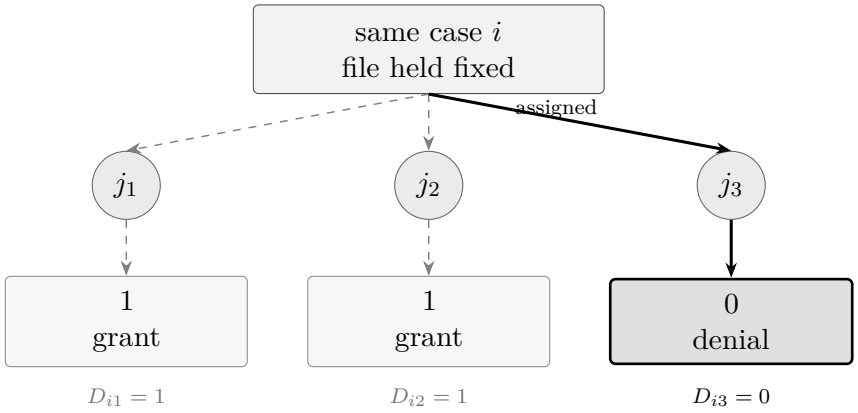
I first give the definition and the measure the dissertation reports, the expected gap between what two officials drawn by chance would decide on the same case. I then address the objections usually raised against the idea: whether an arbitrary decision is the same as a wrong one, which requires discussing what a right decision even is; whether one number can describe an official; whether arbitrariness is just discrimination under another name; and whether removing it always makes decisions better. The chapter extends the theory in *Paper 1*, and it is written to be understandable without the paper.

3.1 The counterfactual object

The dissertation uses the word *arbitrariness* for the counterfactual that a case could have received a different outcome had it received a different bureaucrat. The outcome does not need to be completely random, as if made by rolling a die, and it need not be impossible to explain after the fact. Once the official is known, the decision may follow from that official's views, and the process is still arbitrary if another eligible official would have decided otherwise. The shorter formulation for a binary case could be that an arbitrary outcome "might as well have been the opposite".

To make it clear, we can formalize the problem, inspired by Fischman (2014). We can denote J_i as the pool of officials eligible to decide case i, and let D_{ij} be the decision case i would receive if official $j \in J_i$ decided it. The process is not arbitrary for case i when all officials would decide alike, $D_{ij} = D_{ik}$ for all $j, k \in J_i$, and it is arbitrary to the extent that they would not. For simplicity's sake, we first look at binary decisions, so that $D_{ij} \in \{0, 1\}$. However, we might as well include other positions in between, like a partial grant, which would apply to settings where the outcome is continuous, e.g., monetary benefits within a range.

Imagine three officials are eligible to receive a case. The first two would grant and the third would deny: $D_{i1} = 1$, $D_{i2} = 1$, and $D_{i3} = 0$. Before the actual assignment of an official, the citizen's administrative position is the whole schedule $(1, 1, 0)$. The citizen will, however, only observe one entry based on the assignment of an official, either 1 or 0, so if the assignment goes to the third official, the case is denied. The other two entries are counterfactual for this case. A case in real life is always and only decided once, so it reveals only one entry of the schedule, and the comparison that the definition asks for is therefore never observed directly; this is similar to what Holland (1986) calls the fundamental problem of causal inference.



This formalization can seem a bit redundant, but it highlights a few things about the concept that easily could be misunderstood. First of all, as this is for now a purely theoretical argument, it in principle applies to every setting where there is some assignment of officials; the point of the dissertation applies not only to settings with random allocation of officials. The use of settings with random assignment is for estimation purposes only. Second, we do not need a correct benchmark. For this, in principle, it does not matter whether or not a correct decision would be what the law strictly says, some sort of rule breaking or rule bending (Maynard-Moody and Musheno, 2003; Schmidt et al., 2025), a model of what the median voter would want (Prendergast, 2007), or whatever other interpretation and opinions on what a correct decision of a case should be. When we are interested in this quantity, we do not need to be able to point to specific cases to call out errors, we just need to know which cases would have received a different outcome depending on the draw of official. I expand on this in the next section.

3.2 Benchmarks and wrong decisions

In the last chapter, I showed that there are different takes on the degree to which there actually exist right answers to administrative decisions. To distinguish what is actually bias, what is variance, and what is an error, we can think of many different benchmarks. In the non-exhaustive Table 3, I have listed five takes on what a benchmark could be. I divide them by when, relative to the decision, the benchmark is specified or becomes observable; other classifications may be more useful for comparing the benchmarks themselves.[2] These different benchmarks can overlap, and hopefully they do so in most cases in real life, but they distinguish between different normative and methodological interpretations of how to approach and measure decisions.

The benchmarks differ in where they allow us to locate the standard for an appropriate decision: before the decision (ex ante), after it (ex post), or within it (in itinere). Benchmarks such as the ones that rely on governing legal or administrative standards or ideas of a democratic mandate set the standard outside the official deciding the case and ex ante for the official receiving the case. To be able to rely on such a benchmark would thus require measuring the congruence between the ex ante measure and the decision itself. It is such a benchmark that *Paper 3* relies on when measuring the difference between the first decision and the later review

[2]See, for example, Dahler-Larsen (2019).

Benchmark	What counts as the appropriate decision	Examples from the literature
Ex ante: standards specified before the decision		
Governing legal or administrative standard	The disposition permitted by law, authoritative doctrine, or an authorised administrative rule	Reversal or remand in *Longer Laws, Better Service?*; audited errors in e.g. Young et al. (2024), Greer and Bullock (2018), and Ban and Hill (2025); the state-agent narrative (Maynard-Moody and Musheno, 2003); "the compliance model" in Brodkin (1997)
Democratic mandate	The disposition in line with elected principals or supported by the public	Expressions of elected representatives or principals (Finer, 1941); Friedrich's popular sentiment (Friedrich, 1940); the median voter's preferred allocation in Prendergast (2007)
In itinere: standards specified while the case is being decided		
Professional or technical standard	The disposition supported by accepted knowledge and competent professional practice	Friedrich's "inner-check" (Friedrich, 1940); Mashaw's professional-treatment model (Mashaw, 1983); deliberative routines (Møller, 2021); regulators tailoring penalties to reduce pollution and limit compliance cost (Kang and Silveira, 2021)
Moral standards or citizen needs	The treatment warranted by the citizen's self-defined needs, conduct, or moral claim	Workers helping according to perceived worth (Maynard-Moody and Musheno, 2003); rule-breaking in response to deservingness (Schmidt et al., 2025); Zacka's demands of care and responsiveness (Zacka, 2017); Officials oriented towards citizens' self-defined needs and circumstances (Engster, 2020)
Ex post: criteria for the appropriate decision observed after the decision		
Observable later outcome	The disposition that best predicts or fulfills a specified observable criterion	Pretrial rearrest (Kleinberg et al., 2018; Angelova et al., 2026); missed pneumonia diagnoses (Chan et al., 2022) or other later patient outcomes (Currie and MacLeod, 2017; Currie and MacLeod, 2020); later job performance (Highhouse and Brooks, 2023, p. 520); later abuse-related hospital admissions of children (Grimon and Mills, 2026)

Table 3: Different benchmarks can define different appropriate dispositions for the same case.

by the National Board of Appeals in Denmark.

Alternatively, some studies use different measures ex post to argue whether or not a decision was correct. Those could be measures of how the overall aim of the policy is fulfilled in actuality, disregarding the actual regulation designed to achieve it, or they could be measures of other ex-post-related outcomes at the individual level e.g. in one study, a good decision for whether or not someone should be released pre-trial is measured as whether or not the person reoffends (Kleinberg et al., 2018). These ex post benchmarks thus require data on later behaviour after the decision.

The last two types locate the appropriate decision at the place it is being made. It is not possible to know before the actual citizen is in the office or the actual file is on the desk what the appropriate decision should be. The decision is also not evaluated on whether or not the outcome turned out positive or negative. Instead, an appropriate decision was one that was correctly tailored to standards of the profession or to the citizen's needs.

For a large share of public decisions no benchmark can be observed at all, which is arguably why this dissertation is relevant to begin with. If errors were easy to observe, there would probably be very few of them. Street-level agencies pursue goals that are vague and often in conflict, and their work resists measurement (Lipsky, 2010); in Wilson's (1989) terms, the result of the work often cannot be observed, so there is no later outcome to compare a decision against. Where the right answer cannot be seen, error cannot be measured, and only the lottery between officials can. This is what the rest of the chapter focuses on: a measure of arbitrariness that needs no benchmark, and so still says something about decisions in the many settings where no benchmark can be observed and pinned down.

Deviations from the benchmark

The benchmark distinction mainly shows that there could be many different ways to define an appropriate decision, so one can choose one's own adventure, choose one, and then we denote whatever the appropriate decision is by $\tau(X_i)$ for a case with the facts X_i. This of course assumes that there is in principle only one appropriate decision, even when applying it requires judgment and reasonable officials disagree (Dworkin, 1977, pp. 31–34, 69–71). Whether this is true is a larger philosophical discussion that another person can write a dissertation on.

Using whatever benchmark, we can do some really basic math to show some of the quantities we are interested in. To illustrate, we can hold it up against the schedule of decisions from before. If we take the three officials from before, two of whom would grant and one of whom would deny, and if the benchmark has exactly one right answer, $\tau(X_i)$, we can write p_i for the share of the eligible officials in J_i who would grant; here $p_i = 2/3$.

First off, from this we can know how wrong the pool of officials is. If the right answer is a grant, one official in three would get the case wrong. If the right answer is a denial, two in three would. This is how wrong the pool is on average, how biased away from the correct answer it is. In either case the share of the pool that decides wrongly is the distance between p_i and the right answer, $|p_i - \tau(X_i)|$. However, this number depends on which decision is correct.

The second thing we can know from this is the arbitrariness the citizen faces. What we want

to know is: if we pick two of the eligible officials by chance, what is the chance that they would reach different decisions? This is a key estimand for the dissertation. Writing j and k for two independent draws from the pool J_i, we want to know the expected gap between the citizen's decision and the one they would have received had a different official been assigned the case. This is also called the Gini mean difference, GMD_i.

$$\text{GMD}_i = \mathbb{E}_{j,k} \left| D_{ij} - D_{ik} \right|. \tag{1}$$

For binary cases, the gap between two decisions is simply the chance that two officials drawn by chance disagree, which happens with probability $2p_i(1-p_i)$: one draw yields a granter and the other a denier, in either order. For the example above the numbers are: $p_i = 2/3$ of the pool would grant and $1-p_i = 1/3$ would deny, so two draws disagree with probability $2 \times \frac{2}{3} \times \frac{1}{3} = 4/9$; in a little under half of the possible pairings, the citizen would have received the other decision. The formula stays $2p_i(1-p_i)$ however many officials are in the pool, since p_i is always just the share who would grant. Importantly, for this number, it does not matter which one is right. Averaged over the caseload, $\mathbb{E}_i[\text{GMD}_i]$ is the chance that a citizen drawn at random, by the pure arbitrariness in who decided, would have received a different outcome.

Dividing this gap by twice the mean rate gives the Gini coefficient, $G = \text{GMD}\,/2\bar{r}$, which Anderson et al. (1999) use for inter-judge sentencing disparity and Byrne et al. (2024) for asylum rates. It is the same gap relative to how often the decision is made at all. It can change rankings of arbitrariness across settings, since an office with a low base rate can look unequal on a small base, and a fall in the Gini, which implies a greater equality between officials, could be due only to a rise in the mean rate. I therefore focus on the absolute gap because that is what a citizen faces.

In the example we could see the whole schedule, all three entries, and so we could simply count. However, this we cannot do with real data. Unless we have appeals or audit data, each case is only decided once, so p_i and GMD_i are counterfactual for every real case. To estimate these counterfactuals, the dissertation uses different approaches. The first and easiest one is to give many officials the same file, which is what the conjoint experiments in Study 2 of *Paper 1* do: every respondent makes the same comparison, so similar to the fictitious example above, the schedule is observed and GMD_i can be counted directly, as the share of pairs of respondents who choose differently. The second is to use random assignment. If cases are assigned randomly to officials, or as good as at random conditional on a particular court, office, or year, as in the settings from the leniency designs in Study 1, the cases an official receives are a random sample from the same caseload as every other official's. The grant rate $r_j = \mathbb{E}_i[D_{ij}]$ is therefore an average over the same distribution of cases for every j, which is why any difference between two officials' rates, $\mid r_j - r_k \mid$, is attributable to the officials and not to their cases.

The difference in the rates is, however, only the lower bound on the amount of arbitrariness, since the rate difference does not capture all same-case counterfactual differences (Fischman, 2014). If two officials grant the same share of cases, but they grant different types of cases, those switches would cancel out in the rates. The difference between two officials is therefore the lower bound on how many cases they would decide differently. The number is however exact if the pair of officials is monotone, i.e. all of the cases the strict official would grant, the lenient one

would also grant plus potentially more, so that the two differ only on the cases between their thresholds. Without random assignment, the rate gap mixes the officials' own tendencies with the cases they happened to receive.

3.3 The general case

To get a general understanding of the argument, we do not want to look at only binary decisions. To keep some continuity in the chapter, we stick with the notation from above, although it is slightly different, since it is not restricted to be binary. If we do not restrict it so, we can write the decision more generally as:

$$D_{ij} = \underbrace{\tau(X_i)}_{\text{benchmark}} + \underbrace{b_j(X_i)}_{\text{stable official term}} + \underbrace{\varepsilon_{ij}}_{\text{occasion}}, \qquad \mathbb{E}[\varepsilon_{ij} \mid i, j] = 0. \tag{2}$$

The equation could apply to both continuous ratings or amounts and a latent judgment behind a binary yes-or-no decision.

The term $b_j(X_i)$ is how far official j is from the benchmark $\tau(X_i)$ on a case like this one. It is a stable part, a part that would reappear if the same official decided the same case again. The residual ε_{ij} would be occasion-noise in Kahneman et al. (2021) and is whatever remains once the stable official term has been determined; for example, it captures an official disagreeing with themselves at another point in time through shifts in attention and other things like football games (Eren and Mocan, 2018).

The stable term and the occasion term are easy to tell apart in the equation, but they are hard to separate in real data, since a case is usually decided only once by one official. To separate the two terms, an official has to be seen deciding the same file, or files alike enough to count as the same, more than once. The leniency designs in Study 1 of *Paper 1* have this, because each official decides hundreds of files and so their grant rates average the occasion noise out, meaning that the spread in grant rates is the spread in stable tendencies. In *Paper 2*, there is also only one case per official, so the dispersion across raters is therefore the stable lottery and the occasion noise together, $\mathrm{Var}_j(b_j(X_i)) + \mathrm{Var}(\varepsilon_{ij})$. Being able to separate these two things might be relevant for suggesting interventions; however, for the citizen, it hardly matters.

For now and in the next subsections, for simplicity, we use one number per official for the stable term, meaning that the official has the same stable shift from the benchmark across files. We can think of this as a general tendency toward leniency or strictness. This is the form I use in *Paper 1* to introduce the problem and the form in which leniency designs summarize officials. Its substantive meaning depends on the benchmark. It may describe a tendency to decide more leniently or strictly than the law, a professional standard, or a criterion based on the citizen's later observed behaviour. As a theoretical analogue, Prendergast (2007, pp. 182–184) also represents each bureaucrat by a single stable parameter: the weight placed on client welfare relative to the median citizen's weight. This is not general leniency in his model, but it serves the same simplifying function of representing stable differences among officials with one number.

$$b_j(X_i) = \beta_j \quad \text{for every } X_i, \tag{3}$$

So, how can like cases get unlike outcomes? For a given case, the citizen's decision misses the appropriate decision by $\beta_j + \varepsilon_{ij}$: the official's stable tendency, plus the occasion noise. Across the pool of officials who could have received the case, the stable tendencies have a mean and a spread. The mean, $\bar{\beta} = \mathbb{E}_j[\beta_j]$, is the average bias of the entire pool compared to the appropriate decision. The spread, $\mathrm{Var}_j(\beta_j)$, is how much it matters which official the citizen draws, how much on average an official deviates from the bias of the group, and this is what generates arbitrariness. The occasion noise is $\mathrm{Var}(\varepsilon_{ij})$, and averages to zero for every official.

To spell things out, two questions can be asked of this. The first is how wrong the office is on average. This would just be the bias term:

$$\mathbb{E}_j[D_{ij} - \tau(X_i)] = \bar{\beta}. \tag{4}$$

However, as argued in the previous chapter, the average bias is not what we are interested in. No citizen receives an average; each receives one official. In the average, the official who is too harsh and the one who is too lenient cancel out, even though the two outcomes they decide come out very differently, and a measure that cancels them has thrown away the thing the chapter is about. The second question is therefore how far from the appropriate decision the actual decision is. We therefore look at the squared errors. And because an official's deviation from the pool mean and the day's noise are uncorrelated with each other and with the mean, we can write the basic MSE decomposition of the equation as follows:

$$\mathbb{E}_j\Big[(D_{ij} - \tau(X_i))^2\Big] = \underbrace{\bar{\beta}^2}_{\text{shared bias}} + \underbrace{\mathrm{Var}_j(\beta_j)}_{\text{arbitrariness}} + \underbrace{\mathrm{Var}(\varepsilon_{ij})}_{\text{occasion noise}}. \tag{5}$$

The identity is the familiar bias–variance split in statistics (Geman et al., 1992), and it is close to the idea in Kahneman et al. (2021). What is new here is the administrative meaning of its terms and the public benchmark they are judged against, which is why *Paper 2* introduces it in this framing, where the outcome is a rating and so continuous.

The decomposition also ties back to the two-draw estimand of equation 1. Take the same fixed case i and two officials j and k drawn independently from the pool. The squared gap between their two decisions has expectation

$$\mathbb{E}_{j,k}\Big[(D_{ij} - D_{ik})^2\Big] = 2\left(\mathrm{Var}_j(\beta_j) + \mathrm{Var}(\varepsilon_{ij})\right). \tag{6}$$

The shared bias drops out entirely, because it moves both decisions by the same amount, and what remains is twice the lottery plus twice the noise. This is the formal version of the point made with the binary schedule: comparing a citizen with their counterfactual self does not need any benchmark, and it isolates arbitrariness from the shared error.

These equations are the basics behind the arguments particularly in *Paper 1* and *Paper 2*. In the next sections, I will provide further explanation of adjacent thoughts, extensions and some of the caveats behind this logic.

3.4 Multifaceted bureaucrats

An official can be strict in one corner of the caseload and lenient in another and the one general tendency number per official of equation 3 cannot reflect that. It gives each official the same distance from the benchmark on every file, so an official who is harsh on one kind of case and soft on another is averaged into someone who is neither. To look at that we return to the general form of equation 2, in which the stable term $b_j(X_i)$ depends on the case. We can think of this term, analogous to the earlier term, simply as how far official j would sit from the benchmark, on average over repeated decisions, on a case with the features of case i. We can split this into two parts. The level, ℓ_j, is the official's average distance over the entire caseload,

$$\ell_j = \mathbb{E}_i[b_j(X_i)]\,, \tag{7}$$

This is very similar to what we had before with the general strictness or leniency, which is what the one number β_j was standing in for. The other part was what we assumed away before:

$$\eta_j(X_i) = b_j(X_i) - \ell_j, \tag{8}$$

This is how the particular official deviates from their normal, general strictness when encountering the relevant facts of case i. So to speak, it is the pattern that the official follows (Kahneman et al., 2021). By construction, the pattern averages to zero over each official's caseload, $\mathbb{E}_i[\eta_j(X_i)] = 0$. The scalar model from before is the case in which every pattern is zero, $b_j(X_i) = \ell_j = \beta_j$ for every case. This is of course purely modelling, and it does not necessarily reflect a mental model of officials having a general leniency as a separate view on the particular cases. Leniency is thus an observed decision rate and not a psychological trait.

General leniency, as the first stages measure it, is the caseload-weighted sum of an official's feature-specific tendencies plus whatever constant would be present regardless of the caseload. Kahneman et al. (2021) distinguish between the two as level noise and pattern noise. Two officials can have the same level and still rank cases differently, so that a citizen with one kind of file is better off at the first desk and a citizen with another kind is better off at the second.

What does that mean for arbitrariness? Across a common caseload, pattern differences cannot reduce arbitrariness. For a fixed case i, the official lottery is now the spread of $b_j(X_i) = \ell_j + \eta_j(X_i)$ across the pool, which has three parts:

$$\mathrm{Var}_j(b_j(X_i)) = \underbrace{\mathrm{Var}_j(\ell_j)}_{\text{level}} + \underbrace{\mathrm{Var}_j(\eta_j(X_i))}_{\text{pattern}} + \underbrace{2\ \mathrm{Cov}_j(\ell_j, \eta_j(X_i))}_{\text{alignment}}. \tag{9}$$

The first part is the spread in levels, the lottery the simple model already had. The second is the spread in patterns on this kind of file, which were previously set at zero; since spreads cannot be negative, we have only added dependence on the officials. The third is the covariance, the alignment of the two terms, meaning how well they fit together. It is positive if the officials who are strict in general are also extra strict on cases like case i, and vice versa. We can think of it as whether the officials who are strict in general are also the ones who are extra strict, or unusually soft, on the particular case combination like this one.

There are three things this highlights. First, arbitrariness is understood as more case-specific. Picture two caseworkers at neighbouring desks. One wears jeans with holes in them, has his own band, dreams of producing wine from the vines in his garden, and votes well to the left. The other wears slimmer, stretchy suit pants, cares a lot about the price of wine, and votes well to the right. Hand them a sickness benefit claim with a hospital discharge letter and a clear diagnosis, and they would both grant it. Hand them, however, a claim from someone who turned down a job offer last month because it was a long commute, and the latter official denies while the former grants. Hand them a child-protection case where there are documented injuries, and both would probably look into the case. On the other hand, hand them one that rests on a neighbour's report about shouting, and one of them opens a case while the other does not. Here, the case decides how much the desk matters. The same pool of officials can agree on ordinary files and disagree on files with particular features, and the shared bias $\bar{\beta}$ of equation 4 is case-specific in the same way, so a pool can be right on the average file and wrong at the margins.

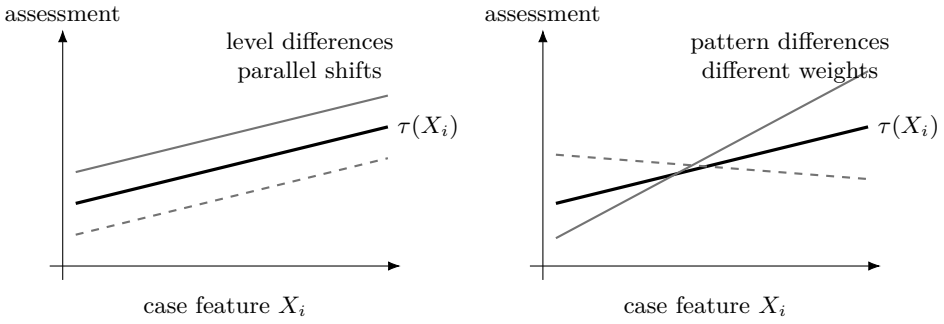


Figure 3: The black line is the benchmark $\tau(X_i)$ and the two grey lines are two officials' stable assessments, $\tau(X_i) + b_j(X_i)$, across a case feature. Left: level differences only; the officials differ by the same amount on every file, so the stricter official is stricter everywhere. Right: pattern differences; the officials weigh the feature differently, so which of them is more lenient depends on the file, and switches in opposite directions cancel in their grant rates.

Second, grant rates that I use in *Paper 1* can hide pattern noise entirely (Sigstad, 2026). Because each official's pattern averages to zero over their own caseload, it is not observable in the grant rates. Two officials with identical rates can disagree on every file, one granting the cases the other denies. Third, a single ranking of officials from strict to lenient is an adequate summary only when the pattern terms are small next to the level terms or line up closely with them.

This highlights the point about the difference in the grant rates being the floor for the amount of arbitrariness. The one number per official understates the size of the lottery. On a single case the third covariance part of equation 9 can have either sign, so for one file the pattern can offset the level, but over the entire caseload it cannot. Because each official's pattern averages to zero over their own files, $\mathbb{E}_i[\eta_j(X_i)] = 0$, the covariance averages to zero as well, and the caseload-average lottery is

$$\mathbb{E}_i[\mathrm{Var}_j(b_j(X_i))] = \mathrm{Var}_j(\ell_j) + \mathbb{E}_i[\mathrm{Var}_j(\eta_j(X_i))] \;\geq\; \mathrm{Var}_j(\ell_j), \tag{10}$$

This is only equal when the pattern part is zero. As such, the lottery a randomly selected citizen faces is at least the lottery the single-number, scalar model reports, and the gap is the average spread in patterns.

There are two designs in *Paper 1*, and these highlight two ways of reaching these parts of the equations. In Study 1, I only measure the levels: a leniency measure is an official's grant rate over a random share of the caseload, so it is ℓ_j on the grant scale, and the spread is $\mathrm{Var}_j(\ell_j)$ on that scale. The conjoint reanalysis in Study 2 is an attempt to estimate the pattern term. Officials there choose between two constructed profiles, so the files are held fixed, and an official's overall tendency therefore simply cancels out, since everyone chooses a profile in each iteration of the conjoints. What is left is how officials weigh the features that differ between the profiles. The paper's finding that most of the stable variation in those weights is official-specific rather than shared is a statement about the second part of equation 9 on contested profiles, and it is invisible by construction to a grant-rate gap. The two numbers cannot be added, since one is a level and the other a pattern on a different scale; together they say that the one number per official is a floor, and that what it leaves out may not be so small.

3.5 Is this not just discrimination?

How is this different from bias? The term "bias" is used for many different things. In this dissertation, I have most often used the term for how a group of people are wrong, e.g. if all street-level bureaucrats are on average wrong about a decision the system would be biased. Prendergast (2007) uses the term for the individual bureaucrat's tendency, so that an individual bureaucrat is biased when they decide on average differently from the group they are in. Often when people use the word in layman's terms, what they mean is whether bureaucrats are more wrong about specific people than others, which would then lead to discrimination. It is therefore worth clarifying what that would look like.

The previous subsection let the stable term vary with every feature of the case. Here I return to the simple, scalar model of equation 3 and add back a single feature, a group feature, which is not included in the benchmark. If officials treat claimants of one group worse than claimants of another, is that not the whole story? Specific groups can be treated differently through the two terms we have set up. We can let G_i denote membership in a group (e.g. gender, race, religion) which is irrelevant to the benchmark, and give the group term its own bureaucrat term:

$$D_{ij} = \tau(X_i) + \beta_j + \underbrace{\delta_j G_i}_{\text{group effect}} + \varepsilon_{ij}, \qquad \delta_j = \bar{\delta} + \widetilde{\delta}_j, \quad \mathbb{E}_j[\widetilde{\delta}_j] = 0. \tag{11}$$

The shared part $\bar{\delta}$ is the average tilt against, or toward, the group, i.e. what every official in the pool adds to a case because of the group the citizen belongs to. That is discrimination as it is usually measured, a coefficient on group membership. The official-specific part $\widetilde{\delta}_j$ is a second lottery, which the citizen also faces in addition to the one described above. This denotes that some officials tilt more than the pool and some less, or the other way around, so that a member of the group faces not only a worse average but also more arbitrariness. For a fixed case i in

the group, the stable official term becomes $\beta_j + \delta_j$, and the official lottery for that group differs from the lottery for citizens outside the group by

$$\mathrm{Var}_j(\beta_j + \delta_j) - \mathrm{Var}_j(\beta_j) = \underbrace{\mathrm{Var}_j(\delta_j)}_{\text{discrimination lottery}} + \underbrace{2\ \mathrm{Cov}_j(\beta_j, \delta_j)}_{\text{alignment}}. \tag{12}$$

$\bar{\delta}$ is discrimination as it is usually measured: the average amount every official adds to, or takes from, a case because the citizen belongs to the group. It could be any shared bias that is attached to a feature of the citizen. $\mathrm{Var}_j(\beta_j)$ is the arbitrariness already defined in the equations above. What is new is $\widetilde{\delta}_j$: how much official j's treatment of the group differs from the pool's average treatment of it. Its spread, $\mathrm{Var}_j(\widetilde{\delta}_j)$, is a second lottery that only members of the group face, and the alignment term in equation 12 says whether it lines up with the first; whether the officials who are strict in general are also the ones who discriminate most. The three can be present in any combination.

This shows that a study that finds no average bias has not shown that group membership does not matter; it has shown that it does not matter in the same direction for every official, and a member of that group still faces a second lottery on top of the first.

Figure 4 shows the two cases with three officials. In both panels the solid line is what each official gives a citizen not in a specific group that is treated differently, and the dashed line is what the same official gives a citizen within a particular group with otherwise the same case. On the left, the officials differ in general strictness; the lines slope, but the gap between the lines is the same for every official, meaning that $\bar{\delta} \neq 0$, but also $\widetilde{\delta}_j = 0$. The citizen in the group is worse off everywhere, but by the same amount, pure discrimination without a lottery.

On the right the officials agree on the other citizen, but disagree about the citizen in the specific group. The first official is harder on the marked citizen; the second treats them alike; the third is easier. The gaps average to zero, so $\bar{\delta} = 0$ and a group coefficient reports no discrimination, while $\mathrm{Var}_j(\widetilde{\delta}_j)$ is large. This is a group lottery without discrimination on average.

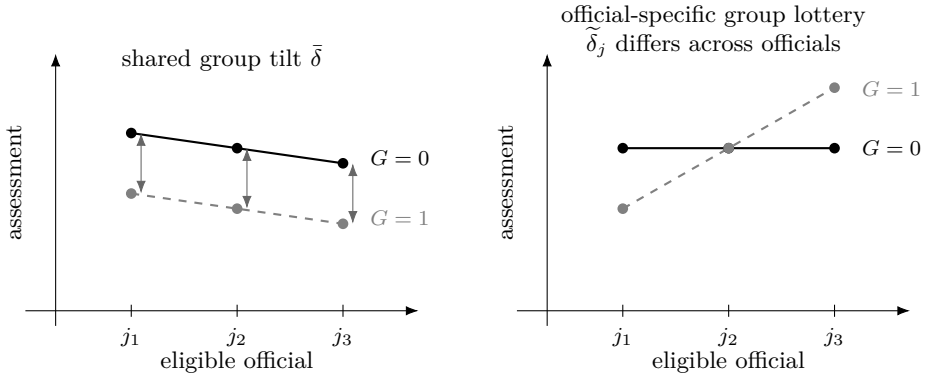


Figure 4: Two ways group membership can enter. Left: officials differ in their general assessments, but the marked group (dashed) is shifted by the same amount for every official, so $\bar{\delta} \neq 0$ and $\widetilde{\delta}_j = 0$. Right: the group shifts are negative, zero, and positive across the three officials; with equal weights they average to zero, so $\bar{\delta} = 0$ while $\widetilde{\delta}_j$ varies. The two can also occur together.

The leniency designs in Study 1 in *Paper 1* will mislabel the discrimination as part of the overall leniency, but will include it. An official's grant rate averages over the whole caseload,

so official-specific discrimination is a part of the general leniency, depending on the size of the group in the data. If the ranking of officials differs by group, however, this will fail the monotonicity assumption and the measure of the spread is thus a lower bound on the total amount of arbitrariness. Frandsen et al. (2023) makes a test for it, which many of the individual studies used in the paper report on. However, the only way to see the group term is to observe the group.

Additional analysis: Comparing discrimination and leniency

The focus of *Paper 1* is to show the importance of arbitrariness. However, for this chapter, to get a sense of how important it is, we can compare it to discrimination by ethnicity, which we know is bad and skews decisions. We can thus ask, what changes a citizen's outcome more, discrimination or arbitrariness from the bureaucrat lottery?

To do so, we need a fitting study that looks at discrimination by ethnicity and for such a question, what better setting than Israel? We can utilize data from Shayo and Zussman (2011), who study Israeli small-claims courts for civil cases. Claims are assigned to judges as if at random within a court. The judges are Arab or Jewish, as are the claimants. Their finding is an in-group bias: a claimant is more likely to win before a judge of their own ethnicity than before a judge of the other. Their estimate is a difference-in-differences, 19.23-percentage-point interaction between an Arab judge and an Arab claimant, meaning how much the gap in win rates between Arab and Jewish claimants changes when the case moves from a Jewish judge to an Arab one, or vice versa.

To compare it with general leniency which we use to identify arbitrariness, we want to know the same quantity: if a claimant's case had been assigned again in the same court, how much would the chance of winning have changed? For the general leniency, it is the expected gap in acceptance rates between two judges drawn at random, the Gini mean difference of equation 1 applied to the judges' rates. In this study that is 4.1 percentage points in these courts.[3] This means that for every 100 citizens in the courts, about four would have won instead of lost, or lost instead of won, had their case been assigned again in the same courthouse. This is disregarding ethnicity.

For the discrimination effect, the answer has two parts. Drawing a new judge changes the ethnic effect only if they are of a different ethnicity from the old judge, which in these courts would happen about a quarter of the time. When this happens, the chance of winning moves by about ten percentage points.[4] On average, ethnicity therefore moves a claimant's chances by about ten points times the one-in-four chance that the new judge is of a different ethnicity, $9.6 \times 0.25 \approx 2.4$ percentage points. This points back to the argument above, which is that a group bias becomes a lottery only to the extent that the draw can change the judge's group: in a court with no Arab judges, an Arab claimant faces the full bias and no lottery at all, the left panel of Figure 4.

Comparing the two numbers, how much general leniency moves the outcome of the citizen

[3] Corrected for judges with few cases, whose raw rates would overstate the spread (see *Paper 1* and Appendix A).

[4] The 19.23-point swing is the sum of the two effects, moving away from an in-group judge to an out-group one and vice versa. $\Delta = 19.23/2 = 9.61$ if the two are equal; see Appendix A.

versus how much the large in-group ethnicity bias moves it, we see that leniency disrupts about 1.7 times as many outcomes, ($4.1/2.4 \approx 1.7$). Neither factor should affect the citizen's outcome; yet the faceless arbitrariness of the plain differences in leniency between judges moves at least as many decisions as ethnic discrimination.

This was mostly meant to be illustrative of the importance of arbitrariness; however, there are, of course, some caveats to this. First, the 2.4 is only the part of the ethnic bias that depends on the draw. Had an Arab citizen been in a court with only Jewish judges, or vice versa, the redraw number would be zero, and the bias would be the estimated 9.6 percentage points. Second, ethically we might care about where the source of injustice comes from; whereas discrimination is also bad for historical reasons, a lottery is less so. Here, we only look at how much it skews decisions on average. Third, this is only one setting; it can be hard to generalise from one case only, which is the aim of calculating meta-estimates as in *Paper 1* and *Paper 4*.

More details and numbers behind this subanalysis can be found in Appendix A.

3.6 Consistently wrong: tails, thresholds, and aggregation

In the model above, arbitrariness always leads to worse decisions, at least decisions away from a benchmark. However, the primary reason to embed this idea of arbitrariness with these different benchmarks is that we of course can imagine when arbitrariness is not all bad. For example, Fischman, 2014 describes two judges who decide every case identically and are always wrong: zero arbitrariness, maximal error. Adding the third judge, a true oracle, would then make arbitrariness greater, but the number of correct decisions would go up. In the binary section, we saw that the error of the case and the arbitrariness in principle cover the same decisions. This subsection examines in which cases the two terms move in opposite directions. I discuss three things: tightening rules, thresholds and decisions in panels.

The bias–variance trade-off

In the next chapter I will discuss different takes on how to reduce arbitrariness. A caveat to doing this, however, is that there is a chance that the intervention could also move the average. We can think about an intervention reducing arbitrariness as potentially restricting or nudging officials into taking particular actions, which would function in the same way as a hard rule. This could mean the officials cannot tailor their decision to fit particular cases, thereby introducing error.

In the equations above, tightening a rule would lower $\mathrm{Var}_j(\beta_j)$, but it might raise $|\bar{\beta}(X_i)|$ on the cases that the rule misfits, and whether the error of equation 5 falls depends on which of these two changes is larger. This is analogous to the bias–variance trade-off in statistics and machine learning (Geman et al., 1992). The decomposition in Equation 5 is the same. A rigid model underfits the data, meaning that it does not catch every nuance, whereas a flexible, variable model overfits, meaning that it interprets noise as signal and bends too much in the direction of whatever data it happened to learn from. We can think of a bureaucrat with a lot

of discretion as a flexible model. They have learned from their own small sample of cases and their personal beliefs, and use those few personal interactions as signals of features that might be irrelevant. Because officials are different and have seen only their own cases, they disagree with each other. Their decision process is so to speak overfitted to their own life.

The trade-off has been measured for a long time and applied to decision making. Meehl (1954) was the first to compare expert judgment with simple formulas on the same prediction tasks, arguing that formulas will typically outperform the expert judgments. Gigerenzer and Brighton (2009) describe judgment as having lower bias, since an expert can respond to features no formula includes, and far higher variance, since experts differ from each other and from themselves; on most tasks the variance dominates, so the more crude rule wins. These are foundational studies behind Kahneman et al. (2021).

Sentencing guidelines in the United States are a classic example. They narrowed the spread between judges, which was their purpose, but critics argued that the guideline sentences themselves were too harsh. What the correct decision is, of course, depends on the benchmarks from before, but in the terms used here: the variance fell, but the centre moved (Fischman and Schanzenbach, 2012; Bushway et al., 2012).

We can use *Paper 3* to theorize how the errors were lowered through the trade-off: A more determinate legal provision narrows what caseworkers can decide, and the finding that longer laws are associated with lower reviewed error rates suggests that, at least in Danish social law over these years, the fall in spread outweighed whatever poor fits between the case and the law the details added.

This will of course not always be true: when the caseload is varied enough that a rule misfits many cases, the bias term will grow faster than the variance term shrinks. A bad rule is still a bad rule, also when it is strict. The next subsection shows that even with the centre held fixed, less spread can mean more error.

Thresholds

Most public decisions are binary, however. As noted, we do not need the decomposition of bias and variance when we have the error term, since the two terms both reflect how many errors there are in a system. However, we can think of underlying measures leading to the binary decisions. This has some implications for whether or not to reduce arbitrariness through lower variance between bureaucrats.

Imagine a municipality that is, on average, stricter than the law when deciding disability-benefit claims. Maybe it is a Danish municipality both under heavy legal pressure to provide services and controlled tightly by a strict budgetary regime.[5] A claimant with a given set of medical conditions is entitled to the benefit, but the particular municipality's shared reading of "reduced work capacity" sets the bar a notch higher, so the typical caseworker denies the claim. The claimants who get the benefit are the ones who happen to draw a caseworker in the lenient tail of the office. Now the office introduces a common assessment guideline to align the

[5] I provided the background analyses for a book under the Danish Power Inquiry on specifically this topic. See Skytte et al. (forthcoming).

caseworkers around the municipality's preferred decision, and the spread between caseworkers indeed falls. The lenient tail disappears and the denial becomes unanimous. Fewer decisions look arbitrary, but more claimants are wrongly denied.

Or take assessors rating an applicant on a scale from 0 to 100, as in *Paper 2*, but with a shortlist threshold at 60. The assessors both rate the applicant and get advice from an AI. This is probably a highly likely scenario. The applicant is in fact a good fit, but the algorithm's score is 55, because the model underweights something the applicant has. Before advice the assessors are spread around 58, some above 60 and some below, and the applicant is shortlisted by a minority of them. After advice the assessors move toward 55 and the spread shrinks. The rating of the applicant is thus less arbitrary, but the applicant is now rejected by all of them. The intervention turned an arbitrary right answer into a consistent wrong one, and a study that measured only the spread would record it as an improvement. This is similar to Cowgill (2019, pp. 10–11), who makes the same point for algorithms trained on human screening decisions: a biased screening process can be corrected only if the humans in it were noisy enough for the right candidates to sometimes get through, and where they were not, "the reduction in noise will actually make outcomes worse."

We can easily apply this to the models, if we think of equation 5 and interpret it as an underlying assessment and a threshold on top of it:

$$M_{ij} = \tau(X_i) + \beta_j + \varepsilon_{ij}, \qquad D_{ij} = \mathbf{1}\{M_{ij} \geq \kappa\}. \tag{13}$$

M_{ij} is how strong a case official j thinks case i is, and the case is granted when that assessment clears the bar κ. For a fixed case i, the pool's assessments are spread around their centre $\tau(X_i) + \bar{\beta}$ with a total spread of $\sigma^2 = \mathrm{Var}_j(\beta_j) + \mathrm{Var}(\varepsilon_{ij})$, the lottery and the noise together. The share of the pool that grants, p_i, is the share of assessments above κ. If we assume a normal distribution, it is

$$p_i = 1 - \Phi\left(\frac{\kappa - \tau(X_i) - \bar{\beta}}{\sigma}\right). \tag{14}$$

Suppose the right answer is a grant, so $\tau(X_i) \geq \kappa$. Whether or not arbitrariness is good or bad depends on which side of the threshold the pool's average is. See Figure 5 for the two cases. In each panel the curve is the whole pool of officials spread around their centre, the dashed line is the bar, and the share of the curve on the wrong side of the bar is the error rate. The dotted curve is the same pool after an intervention that narrows the spread without moving the centre; it has the same area, since it holds the same officials, so it is taller and thinner. The shaded part of each curve is the share of officials who decide rightly, grey before the intervention and hatched after; what the intervention does to error is the change in that share. If the centre is on the right side, $\tau(X_i) + \bar{\beta} \geq \kappa$, the officials who deny are the ones in the lower tail, and spread is what puts them there. If we shrink the arbitrariness, σ, toward zero and p_i moves to one, every official grants, and the error disappears. If the centre is on the wrong side, $\tau(X_i) + \bar{\beta} < \kappa$, the officials who grant are the ones in the upper tail, and spread is the only reason there are any who correctly grant. If we shrink σ toward zero and p_i goes to zero, every official denies, and everyone is wrong.

Each of the two panels is one type of case. A full caseload would be a row of such panels,

and the same pool is on the right side of the bar for some files and on the wrong side for others. The same reduction in arbitrariness lowers the error rate on the cases the pool gets right on average and raises it on the cases the pool gets wrong on average. On the underlying assessment scale however, equation 5 is still the same and the lottery is still a pure loss compared to the benchmark; it is only once we introduce the threshold that it becomes, in the decision the citizen sees, a loss on one side of the bar and a gain on the other.

This is also called zero-one loss of the bias–variance split. Friedman (1997, pp. 61–62) calls the sign of the centre's position "the boundary bias". We can summarise this as follows: with zero variance the decision is always right when the centre sits on the right side and always wrong when it sits on the wrong side.

This highlights two things for the use of tools to reduce arbitrariness in bureaucratic decisions, as I will discuss later in the next chapter. First, it is really important that the models are not wrong on average. Second, it is important to know which parameter one is decreasing the loss function for: there was less error in the underlying assessment, but once it was used for something it became harmful. This is why knowing the bias is still important when seeking to reduce arbitrariness. For real-life applications, however, it is probably more likely that the decisions are right on average, but this is a good thing to keep in mind.

Panels

A common idea to reduce variance is to aggregate, mimicking "wisdom of the crowds" arguments. It is also hailed in the public administration literature as a good thing e.g. in arguments in favour of deliberation between street-level bureaucrats (Møller, 2021). However, letting many officials decide together also has some of the same problems that we saw with the thresholds. This is Condorcet's (1785) jury theorem.

Take a case with a correct decision and a panel of n members who vote independently, each right with the same probability q; the panel follows the majority. If q is above one half, the chance that the majority is right rises with n and goes to one as the panel grows. If q is below one half, it falls with n and goes to zero.

A panel is made to reduce the lottery, since it replaces one draw from the pool with an average of several, and like a template or an anchor it removes the lottery in whichever direction the pool already leans. If we go back to the example from the beginning of the chapter, three members drawn from a pool that is right two times in three reach the right answer in 20/27 of cases. That is better than one official. However, three drawn from a pool that is right one time in three reach it in 7/27, which is worse than one.

This is also why, in *Paper 1*, panels are treated as one decision-maker. What the panel does to the individual lottery is known from its rule; what it does to error depends on where the pool's centre sits, and that is fundamentally unknowable.

From these two examples of thresholds and panels, we see that to reduce arbitrariness is not the same as reducing error. The two go together on the cases the pool gets right on average and they go apart on the cases it gets wrong on average. Tools that shrink the spread—including a template, an anchor, a more determinate rule, or a larger panel—change both the arbitrariness

and the error, but we can only make assumptions about the latter.

Normatively however, it depends on the benchmark whether or not we see arbitrariness as intrinsically bad or whether it is only bad because it leads to errors. The right decisions in the tails of these distributions are only received by luck and are thus still arbitrary. The claimant who draws the lenient caseworker in a strict office gets what the law entitles them to, but gets it because of the draw, and could not have counted on it. The citizen's position depends on whose desk the file lands on, and a right answer received that way does not change the dependence on the bureaucrat, as Pettit (1997) criticises. On either account, an intervention that reduces arbitrariness should be judged on both quantities, the spread it might remove and the new center it might introduce. *Paper 2* measures the first quantity and not the second, and its result should be read with that limit in mind; *Paper 3* observes reviewed error, which is the second quantity, and cannot say whether the determinacy that lowers it did so by moving the centre or by shrinking the spread. Where the centre is wrong, the remedy is to move it, not to tighten the pool around it; if the center is right, removing arbitrariness will be good on all accounts. That is what the next chapter is about.

3.7 Conclusion

This chapter defined arbitrariness to separate it from the things it gets confused with. Arbitrariness is that a fixed case's outcome depends on which official gets it and thus exists in a lottery. We can measure it as the expected gap between what two officials drawn by chance would decide, equation 1. It is not necessarily error, but that is the neat part. A lottery can be measured without a benchmark; error cannot. It is not bias either. Bias is how much the pool is skewed in one direction; the lottery is how much the pool is split, and equation 5 adds them. It is not occasion noise, though one decision per file cannot separate the two. And it is not discrimination: a tilt shared by every desk is a bias, the official-specific part of it is one more lottery. To show the difference, the chapter found that in the Israeli courts the general lottery moved at least as many outcomes as a large ethnic bias did.

The chapter also showed what reducing arbitrariness cannot promise. Whether making officials more alike leads to better decisions depends on whether the officials are right on average. Where they are, the disagreeing few are the ones making mistakes, and bringing them into line removes those mistakes. Where the officials are wrong on average, the disagreeing few are the only ones getting it right, and bringing them into line means everyone is wrong, consistently. The usual tools cannot tell these two situations apart: a shared template, an algorithm's recommendation, a larger panel, or a more detailed rule all push officials toward whatever the majority thinks.

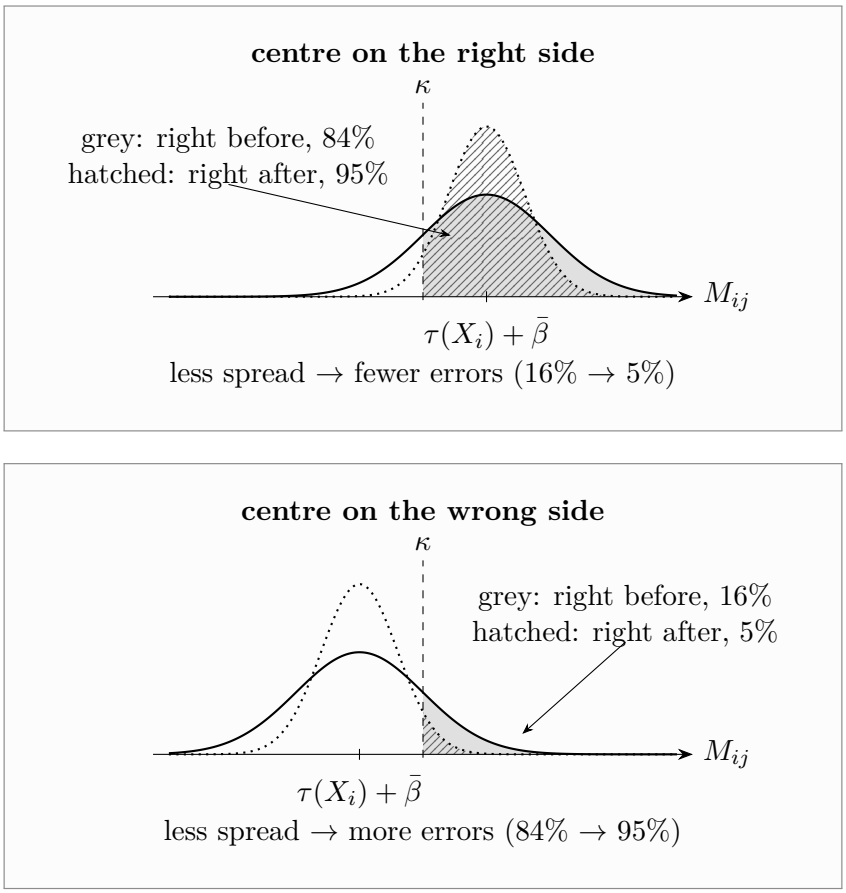


Figure 5: One fixed case, the right answer a grant, and the spread of the pool's assessments M_{ij} around the threshold κ. Each curve is the whole pool, so the two curves have the same area, the numbers are made up and depend on the shape of the distribution. The solid curve is the pool as it is, the dotted curve the same pool after an intervention that cuts the spread to six tenths without moving the centre $\tau(X_i) + \bar{\beta}$. In both panels the shaded area is the share of officials who decide rightly, that is, who grant: grey before the intervention, hatched after. Top: the centre is on the right side of the bar, so most officials already grant, and narrowing the spread raises the share from 84 to 95 percent; error falls from 16 to 5 percent. Bottom: the centre is on the wrong side, so only the upper tail grants, and narrowing the spread cuts that tail from 16 to 5 percent; error rises from 84 to 95 percent. The same intervention on the same officials, and the opposite effect on error.









Chapter 4

How to Get Like Cases Treated Alike

In the previous chapters, we have pointed to the problem of discretion which is that it can make like cases not receive like outcomes. This problem is similar in the take-up literature, which however focuses on the citizen side of the bureaucratic encounter. Whereas in the last chapter, I discussed abstract ways in which systems can be designed to lower this rate, I will now turn to some actual policy interventions that could be made. Using Finer's terms, the question is thus how we can increase administrative responsibility, not necessarily administrative efficiency, i.e. interventions that help citizens receive what the law promises, rather than whether administration becomes cheaper or faster (Finer, 1941, pp. 335–336).

The chapter draws on the previous ones. The highlighted interventions broadly follow the two positions from the debate between Finer (1941) and Friedrich (1940) in section 2: an intervention can either reduce the amount of discretion in the rules themselves, so there is less to decide; or an intervention can serve to support judgment where the rule stays more open. I describe them according to which term they reduce in the equations in the previous chapter: the total amount of error directly, systematic bias or variance.

I argue that when judging interventions of administrative responsibility, we should look at the consequences on both the citizen and the bureaucratic side. We can therefore group interventions broadly in a two-by-two matrix as either being (1) Finerian and rule-based or (2) Friedrichian and support-based; as well as either (a) targeting the citizen or (b) the bureaucrat margin.

I discuss the interventions at three stages: the stage of creation of rules, the interpretation and judgment around rules, and citizens' reactions to rules. The chapter functions as a synthesis for the frame of the dissertation. It serves to tie together the four papers, so we can more broadly understand how to get like cases treated alike. I therefore start by describing why the citizen is important in the overarching question of the dissertation.

4.1 On the citizen's role

Everything in the previous chapter conditions on a case reaching a desk. The counterfactual that defines arbitrariness compares the officials who could have received a case. For the eligible citizen who never learns of a benefit, never applies, or gives up halfway through the forms, there is no official to compare. So, if the goal is to have perfect "congruence" between policy as stated and policy as implemented (Fuller, 1964), it is not enough to optimize how the bureaucrats handle cases. Thus, like cases risk not receiving like outcomes before the bureaucrat lottery is even drawn.

Of course, these types of failures are not random, or quasi-random as the draw of officials

can be. For most public services and benefits, there is an application procedure in which citizens themselves have to initiate the encounter with the bureaucracy and fill out forms and such. In a Danish context one can imagine benefits such as housing benefits, tax deductions, subsidised day-care fees, and more, even though many of the public services are universal. These experiences of hassle are often referred to as administrative burdens (Herd and Moynihan, 2018; Halling and Bækgaard, 2024). The demands asked of citizens are usually defended as verification, but they also work as a second eligibility test that the law never specifies: a test of reading skill, stamina, paperwork, and time. Citizens who fail this second test do not take up the services and benefits they are entitled to (Oorschot, 1991; Ko and Moffitt, 2024).

Burdens are studied as "politics by other means" in the sense that they actually serve the purpose of the politicians who voted for them if they are interested in deterring take-up of benefits and services (Herd and Moynihan, 2018). As such, Nichols and Zeckhauser (1982) describes the demands placed on citizens as ordeals that serve as a means of efficient distribution of services; the citizens who are mostly in need of the benefits and services will be more likely to pay the administrative costs of the ordeals and as such ordeals help distribute the government's resources most efficiently. This theory is however mostly about administrative efficiency rather than responsibility, and even so the core of the ordeals-argument is that the variable moving take-up ideally should be personal utility of the citizen rather than their skill or knowledge.

Furthermore, when asking citizens directly about the trade-off between more burdens and more verification, they lean toward removing burdens rather than keeping them, most clearly in welfare domains, though the preference is weak (Stenderup and Pedersen, 2025). Public opinion, at least in Denmark, is therefore not the main obstacle to removing demands.

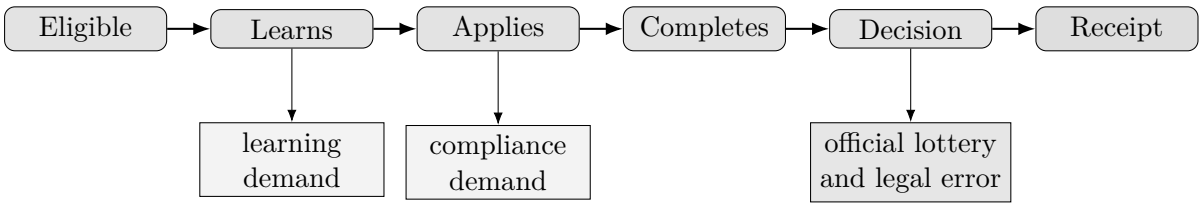


Figure 6: The promise of the law can fail before a decision. *Paper 4* estimates interventions on the early links. The other three papers condition on a case reaching the decision link.

Non-take-up compounds along a chain from when the citizen is eligible in law until they actually receive a benefit. The citizen must learn enough to act, apply, complete what the process demands, receive a decision, and finally receive the benefit. This is visualized in Figure 6. Writing it this way shows why interventions at different links cannot be expected to raise take-up equally: telling people about a benefit helps at the first link and leaves the later ones standing, while removing a compliance demand takes a link out of the chain for everyone. Only after the citizen completes these links do they experience the official lottery, and potentially a legal error.

This extension is important for two reasons. First, in this view, absence is itself a failure, since a citizen who never reaches the decision stage has lost the entitlement just as one who is wrongly refused. Second, participation also changes the decisions: the spread and the centre are computed over the cases that arrive, so a reform that brings new claimants to the desk changes the caseload over which every official-side quantity is measured. The margins are thus causally

connected.

When we are judging whether or not a reform of the rules of the state is a good or a bad thing for administrative responsibility and for like cases to be treated alike, we should thus look both at whether it makes citizens' access to the bureaucracy easier and whether or not it makes decision-making more consistent and correct.

4.2 Two sides of simplification

When using the word "bureaucracy," what most people think of is too many Kafkaesque rules that clutter the public sector and make engaging with it burdensome and tiring. A common political claim is therefore that the law is too long and too complicated, and the remedy is fewer and simpler rules. Nevertheless, generally around the world, legal codes have never been longer (Katz et al., 2020; Jakobsen and Mortensen, 2015; Hinterleitner et al., 2024; Adam et al., 2019; Bendtsen, 2026b). Recently, the most prominent statement is the Draghi report on European competitiveness, which treats accumulated regulation as a burden and recommends cutting it back (Draghi, 2024). Also in Denmark, even across parties, it seems to be a proud tradition for politicians to promise to remove rules. In section B, I have extended the table from Jakobsen and Mortensen (2014, p. 132) to highlight just how often it appears on the political agenda.

Such claims of "the law is too long and complicated" point at two different things. The first is the rule text the official applies: the conditions the caseworker must establish before deciding. The second is the process the citizen must complete: what the applicant must know, document, and do before a decision is made.

From the dissertation there is some evidence on both sides of this question. On the rule side, *Paper 3* finds that longer and more cross-referenced provisions of Danish social law are associated with fewer decisions overturned on appeal. On the process side, *Paper 4* finds that heavier demands on applicants, across many countries and benefit types, lower take-up.

Looking at those two papers only, to increase administrative responsibility, a rule text should state more exactly what the official must establish, and the process should demand less of the citizen. These two recommendations can coexist, if the length of a statute is not a burden on the citizen. However, theoretically, there could be a path from the text to the citizen that runs through the learning demand, since a citizen must understand enough of the rule to know that they are eligible, so a more determinate text could raise the learning cost while it lowers the official's error.

To test those recommendations, I will now go through prior studies on these questions.

4.3 The official's rule

Similar to the popular critique of "bureaucracy", there is also in the political science and public administration literature an argument that the growth of regulations harms implementation of policy. Adam et al. (2019, pp. 97–98) argue that policy growth gives agencies more tasks and larger caseloads while making rules harder to coordinate. When staff and budgets do not

increase, agencies must ration their attention. Knill et al. (2024) call this policy triage: agencies carry out one duty by delaying or neglecting another. Since all formal duties remain in force, administrators in effect decide which political promises they should count first, while policy makers and implementers can each blame the other for the neglected duties. This is where their argument is relevant to arbitrariness. An unwritten or shifting priority can make two like cases receive unlike attention and leave citizens unsure who is responsible.

Triage does not by itself tell us whether implementation is arbitrary. An agency can announce which duties take priority and apply that order to every case. Alternatively, each office can decide quietly which duties it will neglect. Both responses count as triage in this literature, but only the second creates a clear risk that treatment depends on which office receives the case. For this, Fernández-i-Marín et al. (2024) argue that closer links between policy makers and implementers reduce the gap between assigned tasks and available capacity, and Knill et al. (2024) find less triage when agencies can obtain resources or reorganise their work. These studies support the claim that overload forces agencies to choose among their duties. They do not examine whether agencies announce those choices or apply them consistently across offices. The argument therefore links policy accumulation to selective implementation, which is troubling in a classical Finerian understanding, but not necessarily to arbitrary implementation.

The authors argue for a narrower remedy than the typical political calls for simplification suggest. Adam et al. (2019, pp. 49–50) regard blanket deregulation and policies such as "one-in-one-out" rules for the legislation as too crude. Fernández-i-Marín et al. (2024) instead urge governments to match new tasks with added capacity, and thus do not encourage governments to simplify rules to get better implementation.

Then conversely, why might longer laws reduce arbitrariness? Length cannot do so by itself. Extra words may state more exactly what an official must establish, leaving fewer questions on which officials can differ, or simply explain conditions already in the law more clearly, helping officials understand and apply them in the same way. There is some experimental evidence on the second possibility. Charrow and Charrow (1979) and Diamond and Levi (1996) test this by rewriting jury instructions while preserving their legal content. Participants understood the clearer versions better, and Diamond and Levi also found more accurate application of the legal standard. These mechanisms are consistent with the result in *Paper 3*, where a coding of the changes in the law also showed that the changes were by far more procedural than substantive.

The direction of this result was not given in advance, however, since much of the evidence is split. Evidence from the American sentencing guidelines shows that introducing these narrowed the spread between judges, which was their purpose, while critics argued that the guideline sentences themselves were too harsh (Anderson et al., 1999; Fischman and Schanzenbach, 2012; Bushway et al., 2012). When the Supreme Court made the guidelines advisory in 2005, the differences between judges grew again, so the narrowing followed the bindingness of the guidance (Yang, 2014). For inspections, the evidence points the other way: nursing-home inspectors checking compliance with over a thousand regulations agreed with each other far less than inspectors rating only thirty-one much broader standards (Braithwaite and Braithwaite, 1995), and restaurant scores produced under more detailed grading criteria were less consistent (Ho, 2012).

This is not an argument for unlimited rule growth. A rule replaces judgment with a fixed description of cases, and that description should then cover what the law is trying to achieve. In some cases the two come apart: a case fits the description although the law's purpose points the other way (Schauer, 1991). If the differences between bureaucrats are deliberate tailoring to cases, a uniform rule removes the tailoring together with the differences, and decisions can thus become more consistent and worse at the same time (Kang and Silveira, 2021). Interventions that increase the specificity of rules, lowering discretion, therefore follow the argument in the previous chapter about the trade-off between bias and variance. Consistency might not be a goal in itself when we care about decisions being made correctly; however, most often it is.

4.4 The judgment around the rule

There is a wide range of interventions that theoretically could help align judgments around a rule. These could include more training before officials face actual cases; peer review, monitoring, disclosure, and pay-for-performance while they decide them; and audits and internal appeals afterwards. Ho and Sherman (2017) provide a broader review of such interventions. Here, I will highlight only the most relevant ones for the dissertation.

One field experiment comes close to testing Finer against Friedrich directly. In 250 Danish schools, teachers received support for implementing a new reading programme (Hjortskov et al., 2026). Of two treatment groups, one was told to use the material as designed; the other was nudged to use their discretion more and use their own judgment. Teachers rated the programme as more workable and a better fit under the control instruction than under either alternative, and committed more strongly to it than the baseline teachers (Hjortskov et al., 2026, pp. 182–183). Families, however, used the programme more when their teachers had been told to use their judgment than when told to follow the design, and signed up more than the baseline group, though these effects were small and only borderline significant (Hjortskov et al., 2026, pp. 180–181). This experiment thus measured the effects of an intervention aimed at officials by looking at the citizens' experience, showing that these two things can disagree. The version the teachers preferred was the one the families responded less to. The study of course reported average effects.

Peer review could also function to align officials around a rule. In Ho (2017), food-safety inspectors in a government agency were randomized into weekly peer review. Their inspections became more accurate and more consistent, and the gains held when the inspectors worked alone.

It is perhaps this type of intervention that is most relevant for the use of AI and algorithms. Advisory decision support could come from superiors, peers, AI, or algorithms. This is what *Paper 2* tests. It is important for these types of interventions to work that the signal the advice provides is informative; however, this seems like a smaller and smaller hurdle for technology to cross at the current moment.

In this space, Grimon and Mills (2026) looked at algorithms in child protection policy and measured the later consequences for the children involved. The experiment tested access to an

advisory algorithmic risk score for a randomized set of households. The teams did not investigate more children, but the children that were in the screened households were later admitted to hospital for harm less often. The effects were small, however, as expected with such a noisy outcome. The trial is sort of the reverse of *Paper 2*: it measures the average of a later outcome and says nothing about the variance between the workers, while *Paper 2* measures the variance.

There are three cautions about this. First, officials might follow advice selectively. Alon-Barkat and Busuioc (2023) and Selten et al. (2023) find in experiments that decision-makers weigh a signal more when it points in the direction they were already biased towards, which simply reinforced rather than ameliorated the priors they held. Second, effects can fade. Stevenson and Doleac (2024) show that when Virginia attached an advisory risk score to sentencing, judges responded at first, but the measurable signs of use had largely faded within a decade. Third, officials can also override a good signal. Angelova et al. (2026) find that most bail judges did worse than the algorithmic counterfactual when they overrode it, and Hoffman et al. (2018) find that managers who most often hired against the results of a standardised test ended up with worse hires.

Making the signal designed to support judges binding is not a stronger version of the same intervention; it is the same as a rule and the caveats of the rule section above apply to it. This is what distinguishes two modes of AI collaboration in *Paper 2* and Keppeler et al. (2025).

4.5 The citizen's process

The entire point of *Paper 4* is to compare different interventions to increase take-up, although there are many studies measuring non-monetary services and benefits that are equally if not more relevant. The paper also cannot assess features that too few papers have studied, but that might have consistent effects, such as stigma reduction (Lasky-Fink and Linos, 2024). A quick summary for the purposes of this chapter: Because nearly identical experiments have been run so many times across different benefits and institutional settings, *Paper 4* shows that broadly, although with a large degree of heterogeneity, helping citizens on their way to apply for services and benefits has positive effects. That is not so shocking, but the paper also serves to illustrate where in the process citizens drop out; since they can fall off anywhere in the chain of Figure 6, it is harder to get citizens to actually receive the benefit rather than simply apply for it.

The strongest intervention to tackle take-up is to automate service and benefit provision. Where the state already holds the data that settle eligibility in its registers, the benefit can be granted by default. Denmark is well placed for this and does so to a wide degree, since income, family, and residence data sit in linked registers.

Under automatic provision, the spread between officials is zero by construction, and everything depends on the average center, because an error in the design is repeated in every case the system processes (Bovens and Zouridis, 2002). Automated administrative systems have excluded citizens in exactly this way, by applying a wrong rule without deviation (Widlak and Peeters, 2020). The design of an automatic benefit should therefore be judged as a rule.

AI, on the other hand, functions more by assisting citizens' judgments and ability to deal with

administrative burdens. Schmitz et al. (2026) document across several countries and settings a sharp rise in citizens sending requests, claims and complaints (such as the ones I am measuring in *Paper 3*). This highlights how interventions helping each side of the counter can affect the other. From before, we could theorize that longer laws in principle could raise the citizen's learning cost. But especially here with the AI example, we see how the citizen side of the counter also affects the bureaucrat side. A lighter process increases and changes which cases are decided and which are contested.

The argument of Schmitz et al. (2026) concerns administrative efficiency; requests arrive faster than offices can handle them, backlogs grow, and statutory deadlines pass. This framing, however, never counts the rise itself as a gain for administrative responsibility and upholding principles of the rule of law. In Pettit's (1997) argument, what makes a public power non-arbitrary is the ability to contest decisions; if people de facto do not have that ability due to difficulties in dealing with the learning, compliance and psychological costs, that power is arbitrary, also in Pettit's understanding. If AI can help overcome that, which Schmitz et al. (2026) suggests, then the "agentic flooding" they are talking about is beneficial when seen as reducing arbitrariness, especially if governments can increase their capacity to deal with the rise in complaints. As such, the caricature of two AIs talking to each other under the pretence of being a citizen and a bureaucrat might look inefficient on average, but it might help some citizens to claim some justice.

4.6 Conclusion

We cannot say beforehand which of the two types of remedies we derived from the 1940 debate, trusting the rules or trusting the judgment, is better in general. If there is a possibility for disagreement between bureaucrats' decisions, i.e., where the law leaves some room for interpretation and where officials differ, it is likely that there will be arbitrariness, as *Paper 1* documents. Interventions to create better decision-making do not matter, however, if the citizen never applies; only interventions reducing burdens do. There are therefore two margins on which policy can fail at the last step.

	Change the rule (Finer)	Support judgment (Friedrich)
Official margin	More determinate provisions; guidelines and templates; binding signals and automation	Peer review; training; advisory signals, incl. official-facing AI with a fixed signal
Citizen margin	Removing compliance demands; automatic provision from registers	Information; application assistance, incl. citizen-facing AI that completes the process

Table 4: The chapter's interventions, mapped by the margin they work on and by whether they change the rule or support the judgment. The two columns generalize Finer's and Friedrich's answers from chapter 2 to both sides of the encounter

Table 4 maps the interventions discussed in this chapter by which of these margins they work on and by whether they change the rule or support the judgment. The dissertation's own studies fit in these cells: *Paper 3* reports one number on the decision margin, looking at changes

to the rules; *Paper 2* reports numbers on the decision margin, theorizing and looking at whether AI and algorithms can support judgment; *Paper 4* looks at the citizen margin.







Chapter 5

Key findings

In this chapter, I will summarise the four papers in the dissertation.

5.1 Paper 1: Arbitrary Bureaucratic Decisions

In this paper, I look at how many decisions are decided by the bureaucrat assigned rather than the facts of the case. Ideally, this number would be zero or at least very low since the allocation of bureaucrats is arbitrary and thus should not have an effect on bureaucratic decisions, no matter whether these concern foster care placement, incarceration, patent granting, or social benefit delivery. As it is inevitable that policies are implemented by street-level bureaucrats, this normative statement becomes an empirical question: Bureaucratic decisions should not function as a lottery, but we do not know the extent to which they actually do. The paper is therefore about the cost of administrative discretion and how it matters for citizens' outcomes.

Prior studies have focused on specific biases, asking whether or not certain groups or case characteristics on average get differential treatment. These studies can rely on manipulating different case characteristics to study the average effects. This paper, however, focuses on the less aggregated estimand of disagreements between bureaucrats, so we need to hold case characteristics constant.

As cases are not alike, it is hard to know which decisions were made due to the facts of the case and which decisions were made due to the allocation of bureaucrats. Moreover, if specific cases go to specific bureaucrats, we cannot just look at the difference between bureaucrats. We therefore need randomization in order to make the claim that the decisions were indeed decided by the bureaucrat and not the case characteristics. If cases are randomly assigned, then bureaucrats' sets of cases are equal in the aggregate by the law of large numbers, and the difference between bureaucrats is therefore unrelated to the differences in cases.

As street-level bureaucrats have discretion in almost all settings where the citizen meets the state, this is a really broad question. It is therefore preferable to have as-if random variation in the allocation of bureaucrats in many different settings. Luckily, there exists a large body of research that relies on this type of variation, and from which it is possible to repurpose their data in order to answer this particular, seemingly unrelated question. These studies are called leniency-IV studies and they use systematic variation in bureaucrats' leniency in order to measure differences in citizens' propensity to receive a binary bureaucrat decision, such as foster care placement et cetera. I therefore have data on 95 different settings of bureaucratic decisions. The papers have different aims in mind; I, however, care only about the underlying data on their independent variable. I thus have data on over 25 million bureaucratic decisions from almost 100,000 bureaucrats, as-if randomly assigned.

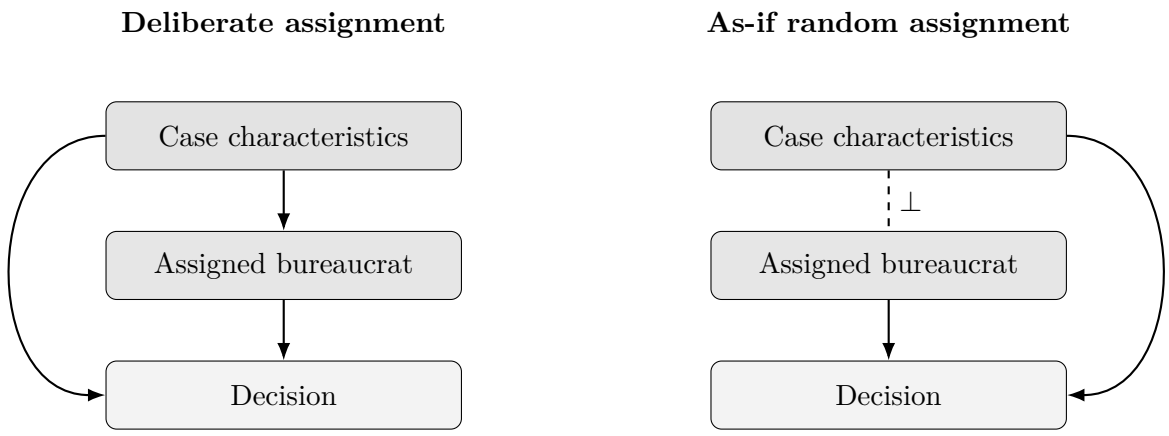


Figure 7: Deliberate and as-if random assignment of cases to bureaucrats. With deliberate assignment, case characteristics affect both which bureaucrat gets the case and the decision, so differences between bureaucrats mix with differences between cases. With as-if random assignment, case characteristics do not affect which bureaucrat is assigned, so differences between bureaucrats show the effect of the bureaucrat.

By assuming that usually more lenient bureaucrats grant every case that usually stricter bureaucrats would grant, and vice versa with denials, the difference between the two bureaucrats' grant rates is how many cases they would decide differently. From this, it follows that some decisions were arbitrary, i.e. might as well have been different, and that some cases were decided solely due to the allocation of the bureaucrat and would have been different if the allocation was redone. It can be thought of as a lottery: some citizens were in the lottery, and some citizens won (or lost depending on the setting).

I get such estimates of arbitrariness in every one of the 95 studies. In the median study, 13% of cases could have been different, had the case been allocated a different bureaucrat. In a quarter of the settings, 29% could have had a different outcome. Two randomly drawn bureaucrats disagree due to difference in leniency in 5.8% of the cases in the median study. As the decisions are binary, if we believe that every case has a right answer, then that means that 3% of cases were decided erroneously. The numbers differ across settings and are larger conditional on whether the citizen received the rarer of the two options.

The numbers do not change much when testing different assumptions about whether or not it is only high-arbitrariness settings that become papers and thus are selected into my data.

In reality, for each of the papers the numbers are too low. This is because we can only identify the disagreements that are systematically due to differences in leniency. If a usually stricter bureaucrat would grant a specific case that a usually more lenient bureaucrat would deny, then that case is not counted as a disagreement. Therefore, there is possibly much more arbitrariness than those numbers suggest.

To look at such disagreements where leniency is constant, I look at how bureaucrats decide differently in experimental settings. Using 13 prior conjoint experiments that mimic actual decisions, real bureaucrats are asked to choose between two different randomized sets of case characteristics. Due to sheer coincidence some of these comparisons will be exactly the same, which means that two bureaucrats have decided the same fictive case twice. When this happens, they agree in the median study 60% of the time. If they chose at random they would agree half

the time, so this suggests that they do disagree a lot.

Credibly and on a large scale, the paper highlights arbitrariness in bureaucratic decisions.

5.2 Paper 2: AI's Role in Enhancing Consistency

Co-authored with Mogens Jin Pedersen and Florian Keppeler

Similar to the first paper, this paper introduces the concept of "Noise" in decision-making into a public administration setting. For the dissertation, the paper's main point is that AI and algorithms can support bureaucrats' judgments and thereby help them make more correct decisions, which also means decisions that are more alike when encountering the same case.

In the paper, we analyse two different collaboration modes of integrating algorithms into decision making. The paper uses data on ratings before and after algorithmic advice from a survey experiment on personnel selection in Germany, but is relevant in multiple other settings where decision-making is based on continuous scoring. The first mode is called ensembling (Choudhary et al., 2025; Keppeler et al., 2025): the final score is a weighted part of the algorithmic advice and the human score. The second mode is called specialization: the human adjusts their score after receiving the advice.

Both types seek to make scoring more consistent: the first does it mechanically, the second behaviourally under full human control. We therefore want to compare these two modes by asking: how much does the behavioural specialization approach reduce noise compared to the mechanical ensembling approach?

We therefore want to find a weighting point where the two modes generate the same reduction in deviations in the scoring, which we in the paper call the "specialization-ensembling weighting point (SEWP)".

In the paper, this number is 13.2%, meaning that giving people an algorithmic score lowers the standard deviation in their scores as much as weighting the final score as 13.2% algorithm and 86.8% human. However, it could easily be extended to other settings where studies are looking at the effect of AI or algorithms on scores.

The mean moved very little in both modes, so in error terms, the advice helped make better decisions on average.

The paper uses algorithmic advice where the variance is zero by definition, but it can easily be extended to accommodate the variance there is in AI, such as large language models. As such, in section C, doctors reacting to AI ratings of cancer risk in Pálfi et al. (2022) have a SEWP of 34%, and predicted recidivism scores in Fogliato et al. (2021) have a SEWP of 12.0%.

The paper highlights how supporting bureaucrats' judgment can help them make better decisions, and this support might very well be from algorithms or AI.

5.3 Paper 3: Longer Laws, Better Service?

Co-authored with Jonas Krogh Madsen

This paper questions whether complex legislation is associated with better or worse decisions.

Since the legal code has grown a lot in almost all advanced democracies, many people advocate for a simplification of laws and regulations.[6] This then invites the question of whether or not that would actually lead to better bureaucratic decisions, or whether it would lead to more errors because the law would inevitably leave more room for discretion. The paper's role in the dissertation is therefore to examine whether stricter rules are associated with better or worse decisions.

Normally, administrative decisions are regulated by many laws, and decisions can thus not be mapped so easily to concrete paragraphs in the law. This is important because just looking at whether or not a particular law became longer and then what happened to the error rates would not be able to separate changes in irrelevant provisions for the decision and would not be able to remove time trends. To look at how the length of the law and the quality of administrative decisions go together, we need decisions that are linked to very fine-grained legal provisions.

In Danish social policy, complaint data are such an example, as they map reviewed decisions to particular paragraphs in the law. The paragraphs grow over time, becoming more specific about procedures and categories through cross-references. They do not change eligibility criteria but provide more details about how the eligible citizens should be treated. As such, we treat complaints as not changing composition, and can therefore treat the aftermath of the complaint as either an error or not.

We therefore have a panel data structure of specific eligibility groups with a governing legal provision that changes specificity over time tied to reviewed administrative decisions. Using a fixed effects model, we find a negative association between the length or the number of cross-references in the law and the rate of reversals and remandings, the error rate.

The paper highlights the importance of thinking about administrative responsibility and correct decision making when suggesting legal simplifications, and for the dissertation it highlights how tighter rules go together with fewer errors.

5.4 Paper 4: Increasing Take-up of Social Benefits

In this paper I meta-analyse a range of field experiments that all seek to do the same thing: raise take-up of a social benefit. In the administrative burden framework, there is a distinction between burden types and what causes citizens to experience the interaction with the government as onerous. In the study of take-up of social benefits, i.e. the problem that citizens do not claim the benefits they are entitled to, this framework is often implied but only sometimes cited. The meta-analysis frames these interventions under the same theoretical framework, in order to distinguish between the types of demands of which the field experiments are actually relieving citizens.

The field experiments are all really similar. Either they provide information about a social benefit or they provide assistance in dealing with the verification process in getting those. Often

[6] For the specific context of Danish regulatory governance of the municipalities, I have made a longer analysis of the specific regulatory tools used to control municipalities in general across over 30 years (Bendtsen, 2026b), arguing that the discretion of the municipalities has decreased. In another paper, I argue that this might be a factor as to why the policy preferences of local politicians might not translate into measurable changes in policy, focusing on gender (Bendtsen et al., 2026).

the studies vary slightly in how the information is framed; however, they essentially seek to do the same thing. The outcomes are also quite similar as they all measure receipt of entitlements that are essentially free money. The difference between the studies is the settings in which they are made, as some study retirement benefits in Finland, and others study tax write-offs in the US. Meta-analyses should always be considered as cross-sectional evidence, and here the analysis provides a great opportunity to see how well findings from field experiments that help citizens carry across settings. One way to do that is through a more narrative review, but the numerical syntheses force comparability more and thus function better to summarise findings across settings.

Although there is great heterogeneity in the effect sizes, most of these studies show positive effects which means that it is fairly easy to help citizens receive their entitlement with a little help, and this is the case across many countries and many different types of benefits. This could also reflect publication bias, so I test for the meta-analytic estimates' sensitivity to publication bias, and find that the overall average effect sizes decrease, but not enough to make them statistically insignificant.

A quick side note. It happens that the results of these sensitivity tests are seen as the "correct" overall estimates, which is an interpretation I would caution against, since all of the correction methods rely on a fairly unrealistic assumption that statistical power in the studies and the true effect sizes should be uncorrelated. This does not make sense since studies do power analyses and think about the size of the effect before doing a study. The best approach is the one by Andrews and Kasy (2019); however, even this relies on that assumption. Publication bias corrections should be seen as sensitivity tests and not as the authoritative source of truth.

The paper finds that effect sizes are largest in the studies that provide assistance versus simple information, although the information ones might be more cost efficient. Similarly, the paper finds that studies measuring on the application stage rather than the actual take-up stage have larger effect sizes. This can show two things. It might show that the experiments potentially target non-eligible people and get them to apply for a benefit they cannot get, ironically placing more administrative burdens on them than before. However, the meta-analysis cannot distinguish this mechanism from another mechanism of later procedural attrition among eligible applicants, which fits the theory of administrative burdens. In either case, the results warn against treating applications as a sufficient measure of take-up.

It is similar in a sense to another meta-analysis in DellaVigna and Linos (2022), however focusing on social benefits across the world seems better to study administrative burdens, since it is not like the outcomes in nudge studies that some people might not want to use, for example large waste item collection, donating to charity or using water filters.

The data for the paper was collected during 2024. Since then, more experiments have been made. There are thus differences in the numbers between the published version in Bendtsen (2026a), and the version in this dissertation. The meta-analysis now encompasses 66 studies with a total of 233 effect sizes. The differences in the estimates are subtle.

Collectively, the paper shows that the problem of lacking take-up is broad, since light interventions across the world can increase it. For the dissertation, it shows that like cases can have unlike outcomes if the citizens are not aware of their entitlements or face many procedural

Estimate	Published	Updated
Overall average effect	6.7 pp	6.0 pp
Actual take-up	4.8 pp	4.4 pp
Application	10.1 pp	9.2 pp
Compliance premium (versus learning)	+7.6 pp	+5.6 pp
Risk ratio, learning demands (actual take-up)	1.24	1.22
Risk ratio, compliance demands (actual take-up)	1.62	1.47

Table 5: Main estimates of Paper 4 in the published version (Bendtsen, 2026a) and in the updated data used in this dissertation.

requirements in applying for them.

5.5 Implications and future research

There are multiple implications to be drawn from the dissertation. The first is apparent. We should study more rigorously whether or not there is large variation in the bureaucracy. It is expensive to have people check procedures and outcomes twice, so if we only rely on that to find errors, we will miss a great deal of the ordinary, invisible injustices that happen more frequently. The dissertation also shows that errors can come not from average biases but from simple variation between bureaucrats. When doing audits, this should be relevant to look for.

Administrative responsibility is not easy to measure. It is easier to get some measure of administrative efficiency, and it is thus easier to prioritize and optimize it. The dissertation has provided some ways of measuring administrative responsibility, which might help guide policy makers to consider broader questions of legality and the rule of law in a more tangible way when considering which effects interventions might have. It is thus incumbent on us to consider how much arbitrariness we believe we can tolerate.

To reduce arbitrariness in bureaucratic decisions, the dissertation suggests two broad ways. First, creating better rules should serve to align bureaucrats not only with their political principals but also with each other, removing banal variation between them that can lead to errors. Furthermore, supporting bureaucrats' judgment and professionalism should equally help them see things eye-to-eye and align their decision-making.

Importantly, ultimately the problem of arbitrariness addresses the problem that what the law says is not how it is being implemented, a problem of congruence. This problem also exists due to factors on the citizen side, and we should not be blind to those. We should therefore seek to simplify processes and ensure that citizens have the right skills and knowledge about their rights.

This furthermore opens up some adjacent directions for future research.

First, we do not know how AI will impact the bureaucracy in the coming years. An obvious research agenda would be to look at whether AI can increase administrative responsibility and support bureaucrats, and perhaps align them without actually removing their discretion and ability to tailor decisions to the individual's circumstances.

Second, we would want to know more about trade-offs. As theorized, stricter rules might

make policies more biased on average, but decrease the variance and as such reduce the amount of errors in bureaucracies. How this could look concretely requires more careful research. Furthermore, questions of trade-offs between other public values and aims are also ambiguous. Interventions that are designed to make the public sector more efficient by making employees "run faster" might end up hurting other public values through introducing arbitrariness. As the studies looking at the predictions of Lipsky (2010) typically focus on average effects, we do not know whether increased coping behaviour also induces arbitrariness.

Third, how do citizens react and think about problems of arbitrariness? For the future, knowing how much citizens actually change their behaviour and their experience of administrative burdens depending on how wide the discretion of the bureaucrats is would be valuable for thinking about how to have a good bureaucracy with a large degree of administrative responsibility. But furthermore, we would like to know what citizens actually think about arbitrariness and what they infer when they themselves encounter errors or suspect that they have done so. For this, it would be particularly interesting to examine the differences in citizens' opinions about errors coming from average biases and errors coming from variance and arbitrariness. Are they perceived to be equally unfair?

Fourth, building on that we would like to know more about where and in which settings arbitrariness is especially likely to occur. The cross-study estimates in *Paper 1* are too noisy to make credible inferences about whether or not certain areas are particularly arbitrary. However, citizens might have different views on where arbitrariness is more and less tolerable and where they would be more inclined to support wider measures of reducing discretion in order to combat it.

In sum, the dissertation provides theory and some answers to the question of how like cases end up receiving unlike outcomes. It suggests ways to measure it, problems in doing so and some interventions to ameliorate it. It thus helps answer a vital democratic problem of how the inevitable discretion of street-level bureaucrats can lead policy away from its intentions, and how to support such bureaucrats to do so less.

Research Articles











Chapter 6

Arbitrary Bureaucratic Decisions

Abstract

Bureaucratic decisions have durable consequences and should not hinge on arbitrary variation. Every case is, however, arbitrarily allocated to a bureaucrat. I therefore seek to measure how many cases are decided by the allocation of the bureaucrat rather than the facts of the case. First, I re-analyse the first-stage results from 88 prior leniency-IV studies across different policy areas and institutional contexts. These studies are not designed to answer my question, but since they exploit variation of as-if randomly allocated bureaucrats, I can repurpose the underlying data. Because assignment is random, the difference in grant rates between bureaucrats is the share of cases that would be decided differently had the case been assigned to the counterfactual bureaucrat. At the median setting, the allocation decides at least 13% of cases. Two randomly drawn bureaucrats disagree on 5.8% of cases at the median; if a case has a correct decision, disagreement implies one was wrong, so for most settings about 3% of decisions is the minimum error rate. These numbers are lower bounds, since they only count disagreements between a usually more lenient bureaucrat and a usually stricter one. Second, to test whether the heterogeneity I observe reflects genuine preference disagreement, I re-analyse 13 existing conjoint experiments using bureaucrats as respondents. In the median study there is only a 59% pairwise agreement among cases with identical observable profiles. I further decompose variation in their choices and find that heterogeneity between bureaucrats explains more variance than agreement.

6.1 Introduction

Bureaucratic decisions have enormous stakes: they determine who receives social benefits, who is incarcerated, whose firm survives regulatory scrutiny, whose children should be taken into foster-care. Given these consequences, outcomes should not be arbitrary, i.e. they should be determined only by factors related to each case, not by incidental features of the decision process. As such, we would like to think that the decisions made by bureaucrats are fair and consistent.

The Weberian bureaucracy is ideally impersonal and does not follow the particular preferences of individual bureaucrats; however, in practice, it is not possible to ensure that identical cases are treated the same way. Bureaucrats will always have some level of discretion that can lead to variation in decisions. On the positive side, higher discretion can allow bureaucrats to adapt cases better to their particular circumstances. This, however, also means that two identical cases can be decided differently simply because they were randomly assigned to different bureaucrats. I call these decisions arbitrary.

In this paper, I estimate the extent to which arbitrariness matters for bureaucratic decisions. I thus want to extract the share of cases whose decision could have changed if the case had drawn a different bureaucrat. I argue that arbitrariness is both widespread across a large range of institutional contexts and policy areas. In situations where bureaucrats have sufficient discretion and the possibility to disagree, they often do, leading to arbitrary variation in outcomes for citizens.

Fundamentally, since Lipsky (1980), it has been known that bureaucrats exercise discretionary power over outcomes, leading to (unwanted) variation in case handling, but the problem has often been framed differently. Earlier research into the consequences of more or less bureaucratic discretion highlighted that decisions might be biased in one direction or the other, e.g. bureaucrats might prioritize easier clients (Brodkin, 2011), prioritize people deemed more deserving (Jilke and Tummers, 2018), or produce racial (Olsen et al., 2022; Keiser et al., 2004), ideological (Gailmard and Patty, 2007; Spenkuch et al., 2023), and gendered biases (Keiser et al., 2002; Lowande and Proctor, 2020). All of these are trying to estimate average effects across bureaucrats reacting to different types of cases. I turn the tables and instead look at the size of the variation between bureaucrats reacting to identical cases. From the point of view of the citizen, it may matter little whether they are treated differently because of biasing factors or noisy factors if the outcome is the same. If one judge is lenient and another strict for reasons unrelated to case characteristics, the resulting arbitrariness harms citizens regardless of whether it can be traced back to a recognizable group-level pattern.

I am interested in two estimands. First, I want to know on average how many decisions would have been different, if another bureaucrat had gotten the case. Second, I want to know how many could have been so. Luckily, there is a body of research that has already estimated this without even highlighting it: studies using the leniency instrumental variable design (Goldsmith-Pinkham et al., 2026; Chyn et al., 2025), where the instrument is assignment to a more lenient or stricter bureaucrat. The studies use the instrument to estimate the causal effects of bureaucratic decisions on downstream outcomes. I am, however, not interested in the downstream causal effects the papers focus on but only in the data underlying the instrument. For this, I collect

135 such studies and filter out studies that are not binary, do not have enough information on the first stage, or use the same setting as another study, so I do not double count. This leaves a separate analysis sample of 95 estimates from 88 studies. The settings span criminal justice, immigration, disability insurance, child welfare, patent examination, corporate bankruptcy, and other bureaucratic settings.

For each of these studies there are cases that are always-takers, whom even the strictest bureaucrat favors, and never-takers, whom even the most lenient bureaucrat denies. They are by definition not subject to bureaucratic arbitrariness as they would receive the same decision regardless of which bureaucrat they were assigned to. Then there are the compliers, individuals whose decision depends on which bureaucrat they draw. A lenient bureaucrat would approve them; a strict one would deny them. I am therefore interested in the complier share.

As assignment is random, we can think of the allocation of bureaucrats as a lottery. The complier share measures how many people were in the lottery, how many could have had a different outcome. I also want to know how many won the lottery, i.e., how often would a randomly selected different bureaucrat decide the case differently? Averaged over a study's caseload, this chance is the Gini mean difference of the bureaucrats' grant rates. That is a measure of disagreement between random bureaucrats, but since two disagreeing bureaucrats cannot both be right, it also highlights the error rate of cases.

After subtracting sampling noise due to differences in observation across studies, we see that in the median study at least 13.1% of cases are decided by the lottery; the middle half of settings lies between 7.6% and 29.2%. This is the 10th-to-90th-percentile difference and thus not the full complier share, since the tails are sensitive. Looking at the P5–P95 difference, the median is 16.8%, with an interquartile range of 9.8% to 37.4%.

Two randomly drawn bureaucrats disagree on at least 5.8% of cases at the median setting (IQR: 3.3% to 12.9%). This is how many decisions the lottery would change were each case assigned anew. Whether these can be labelled errors depends on whether one believes that there is an appropriate, correct decision for each case or that some cases are genuinely indeterminate (Fischman, 2014). However, if cases have correct decisions, half of those disagreements are errors.

Some decisions are rare, and here arbitrariness matters more. For example, few children are removed to foster care. For these people, the draw matters more, so I ask what share would have received a different decision under a new draw of bureaucrats. In the median setting, 14.4% of the rare decisions would have been different had another bureaucrat handled the case. These are removals, incarcerations, and benefit denials that happened only because of the arbitrary assignment of a bureaucrat. The middle half of settings lies between 6.5% and 27.0%. In five of the 95 settings, more than half of the rare decisions would be different under another draw.

For the summary statistics of the distribution of the studies, the medians and the means, there is a problem of selection. Settings that do not have a study written, either as a publication or a locatable working paper, do not figure in the analysis, and since settings without or with very little variation in the first stage are likely not to be written, the numbers could be inflated. As this critique mimics that of publication bias, I test the sensitivity of the summary statistics to selection and find only marginally lower numbers. Furthermore, estimates from settings that do not have a leniency study are larger than or equivalent to those in the pool of studies.

Importantly, these numbers are lower bounds. If the monotonicity assumption, meaning that a bureaucrat who approves more cases overall would never deny a case that a stricter colleague would approve, does not hold for the cases, there are more arbitrary decisions. I cannot measure this, however, since no case is decided twice.

In Study 2 I turn to experimental settings to assess the underlying mechanism: is there substantial heterogeneity in how bureaucrats evaluate cases? A growing number of studies use conjoint experiments to elicit how bureaucrats weigh different case characteristics when making decisions. The conjoint designs measure responses on a common scale. Rather than reporting subjective ratings where individuals may interpret scale points differently, bureaucrats reveal their priorities through forced choices between case profiles. Thus, I hold the leniency level constant and vary instead how the bureaucrats weigh different case characteristics. Disagreement therefore reflects genuine differences in how bureaucrats prioritize case attributes. These studies typically estimate average marginal component effects but do not examine whether the average masks agreement or heterogeneity among the surveyed bureaucrats; I want to know how different they are.

I obtain replication data from 13 individual experiments from 10 conjoint studies with bureaucrat respondents. First, I look at the pairwise agreement rate between bureaucrats who evaluated the exact same case comparisons. Across the 10 experiments where such shared comparisons exist, the median agreement rate is 59% (mean = 61%), meaning that when two bureaucrats face the identical pair of case profiles, they disagree on which case to prioritize roughly two in five times. Second, inspired by Lauderdale et al. (2018), I decompose the variance in bureaucrats' choices into three components: agreement (shared attribute weights, reflecting the originally estimated AMCEs), heterogeneity (systematic differences in how bureaucrats weigh attributes), and instability (within-respondent noise). Of the attribute-explained variation, heterogeneity accounts for a median of 70% and agreement for 30%. In other words, when case attributes predict bureaucrats' choices, it is predominantly because different bureaucrats weigh those attributes differently, though a non-trivial share reflects shared responses to case characteristics.

Overall, I argue that bureaucratic arbitrariness is an understudied problem in bureaucracies and that it reflects deep, systematic disagreement over how to evaluate cases. Importantly, my argument is not that civil servants, judges, and bureaucrats are incompetent or ill-intentioned. I believe they are, by and large, conscientious professionals doing the best they can under difficult conditions. This rather shows that in facing complex cases and with limited guidance, even well-meaning bureaucrats inevitably reach different conclusions. Documenting this variation can be a first step toward finding out which tools and institutional support might help the most.

6.2 Theory

The Problem of Arbitrary Discretion

The opposite of having the rule of law, A. V. Dicey argued, is to have *"government based on the exercise by persons in authority of wide, arbitrary, or discretionary powers of constraint"*

(cited in Waldron, 2011, p. 10). As such, arbitrary discretion has been a canonical concern in the study of bureaucracy. One of the core reasons why bureaucracy is even valuable is because it disciplines discretion such that like cases should be handled according to general rules rather than according to the idiosyncrasies of the bureaucrat who happens to handle them. Weber (1978) even defined the ideal-typical bureaucracy as one having rule-bound, impersonal decision-making. The bureaucracy therefore aspires to treat like cases alike; otherwise, it would be considered an error.

As Lipsky (1980) argued, however, public policies are what they end up being when citizens encounter them. The implementation of the policy is what the policy is. So, it does not matter if a government issues entitlements if the citizen never gets those entitlements. The government is thus reliant on street-level bureaucrats (e.g. social workers, judges, police officers and teachers) to carry out policies faithfully for democracy to work as intended.

As such, discretion cannot be eliminated. As rules are incomplete and cases differ, street-level bureaucrats must interpret how to apply general rules to specific situations under ambiguous goals, fuzzy boundaries, limited resources, and often only soft evidence (Lipsky, 1980; Zacka, 2017). This is simply a structural feature of policy implementation. The question is therefore not whether discretion exists but when it leads to arbitrariness. Even those who defend wide street-level discretion acknowledge a lower bound; a purely randomized decision process is, as Mashaw (1983, p. 93) writes, "an affront to our sense that each case should be decided on its own merits." If the pure lottery is intolerable, there must be some standard of how cases should be decided.

I therefore look at the case level of how much arbitrariness there is, in order to judge how much discretion is actually turning into a problem of like cases not being treated alike. We can never know the potential outcome of each case, since there is for each case only one bureaucrat and one outcome. However, if like cases are not treated alike, there will necessarily be some disagreements between bureaucrats, and where this happens, the assignment mechanism will operate as a lottery over life outcomes. A clear measurable cost of discretion is therefore the size of that affected share, the fraction of cases whose outcome depends on who decides rather than what the case contains. Measuring its magnitude is therefore a first step toward knowing whether discretion, in a given setting, remains within tolerable bounds.

Administrative errors and arbitrariness

An arbitrary bureaucratic decision can be defined without knowing which decision was right. We just need to know whether or not the case depended on a bureaucrat rather than the facts of the case. We can formalize this. Let J_i be a pool of bureaucrats eligible to decide case i, and let D_{ij} denote the decision made by bureaucrat j. The decision for case i is not arbitrary if all eligible bureaucrats would reach the same outcome, $D_{ij} = D_{ik}$ for all $j, k \in J_i$. It is arbitrary to the extent that they would not.

If we further believe that there exists a true decision, whether this is the decision that perfectly captures the law, the will of the principal, or some other Platonic ideal, we can decompose the potential error of such a general decision (Geman et al., 1992; Gigerenzer and

Brighton, 2009; Kahneman et al., 2021):

$$D_{ij} = \underbrace{\tau(\mathbf{X}_i)}_{\substack{\text{appropriate}\\ \text{decision}}} + \underbrace{\delta_j}_{\substack{\text{bureaucrat}\\ \text{tendency}}} + \underbrace{\epsilon_{ij}}_{\substack{\text{within-bureaucrat}\\ \text{noise}}} \tag{15}$$

where $\mathbf{X}_i$ are all case-relevant variables and $\tau(\mathbf{X}_i)$ is the decision a perfectly informed and consistent application would yield. If the decision is binary, the additive form could be read on a probability scale, with δ_j the shift in bureaucrat j's probability of treating the case. The mean squared error around this benchmark can then be decomposed as:

$$\underbrace{\mathbb{E}\Big[(D_{ij} - \tau(\mathbf{X}_i))^2\Big]}_{\text{total error}} = \underbrace{\big[\mathbb{E}(\delta_j)\big]^2}_{\text{shared bias}} + \underbrace{\text{Var}(\delta_j)}_{\substack{\text{variance}\\ \text{(bureaucrat lottery)}}} + \underbrace{\text{Var}(\epsilon_{ij})}_{\substack{\text{within-bureaucrat}\\ \text{noise}}} \tag{16}$$

Bias captures systematic deviations from the benchmark shared across decision-makers. Variance captures how much decisions shift depending on who decides, holding the case constant. Noise reflects within-decision-maker inconsistency driven by irrelevant contextual factors. In a single decision ϵ_{ij} bundles two components that only repeated measurement can separate: a stable case-by-decision-maker interaction (officials disagreeing on particular kinds of case) and day-to-day occasion noise. Because D_{ij} is a potential decision, this decomposition holds regardless of how cases are allocated; if assignment is non-random, δ_j still exists but is no longer separately identifiable from case composition, so random assignment is a condition for measurement, not for the existence of arbitrariness. Writing it this way, we look at the decision-maker's tendency as additive, a scalar shift, so in this model officials cannot be lenient on one type of case and strict on another; such case-specific tendencies are absorbed into ϵ_{ij}, making $\text{Var}(\delta_j)$ a lower bound on total between-decision-maker disagreement. In Study 1 this is assumed in the partial monotonicity assumptions, where violations attenuate the observed ranges; Study 2 relaxes this restriction.

This decomposition has a simple normative implication: taking the quality of decisions as given, predictability is desirable. Regardless of how one defines the appropriate decision, whether as the principal's policy intention, the median bureaucrat's view, or some Platonic ideal, $\text{Var}(\delta_j)$ is a pure loss as long as we believe that there exists an appropriate decision. For any fixed τ, dispersion among bureaucrats reduces accuracy without compensating benefit. This is a property of the decomposition itself; thus, it does not require taking a position on what τ should be. We do not need to know the correct decision, just acknowledge that there is one, to say that variance is bad.

Similarly, if $\text{Var}(\delta_j)$ can be reduced without altering τ, then that would only be beneficial. If reducing it requires changing τ as well, for instance by constraining discretion in ways that also limit the institution's ability to tailor decisions to case characteristics, there is a bias-variance trade-off. The point here is that $\text{Var}(\delta_j)$ is an important standalone quantity because it represents a component of error that is undesirable under any specification of τ.

There are a few studies highlighting and measuring arbitrariness in implementation. Already, in 1933, Gaudet et al. (1933) showed that among thousands of cases heard by six New Jersey judges, the share of defendants sent to prison ranged from 34 to 58 percent. Also, Mashaw

(1983, p. 87) on denial rates in the Social Security disability system, Keiser and Soss (1998) in one study of the good cause exemption in U.S. child support enforcement, and a line of studies on asylum judges (Ramji-Nogales et al., 2007; Rehaag, 2008; Byrne et al., 2024) all showed large arbitrariness, with ranges from 60 to 90 percent. Best et al. (2023), using a variance decomposition of Russian public procurement decisions, find that 39% of quality variation is attributable to bureaucrat heterogeneity.

The literature on policy implementation, discretion, and street-level bureaucracy has, however, primarily focused on the bias and the within-bureaucrat noise terms. Research on bias has studied systematic differences in how identifiable groups are treated: racial discrimination in administrative access and responsiveness (Einstein and Glick, 2017; Keiser et al., 2004; Olsen et al., 2022), cream-skimming and deservingness judgments (Brodkin, 2011; Jilke and Tummers, 2018), and ideological or gendered differences in implementation (Gailmard and Patty, 2007; Spenkuch et al., 2023; Keiser et al., 2002; Lowande and Proctor, 2020). These are questions about average tendencies: whether decisions systematically move in one direction for identifiable citizens, cases, or bureaucratic contexts.

Research on within-bureaucrat arbitrariness, meanwhile, has studied the temperature outside the courthouse (Heyes and Saberian, 2019), the outcome of a college football game (Eren and Mocan, 2018), and the time since a judge's last meal break (Danziger et al., 2011), methodological controversies aside. This work documents that $\text{Var}(\epsilon_{ij})$ is non-trivially large, that is, that decisions are swayed by irrelevant circumstances.

What this framing misses is that from the citizen's perspective, both $\text{Var}(\delta_j)$ and $\text{Var}(\epsilon_{ij})$ function as randomness. The citizen does not know whether a harsh outcome reflects an unlucky examiner draw or the examiner having a bad day. This is why Kahneman et al. (2021) labels both terms "noise." Yet $\text{Var}(\delta_j)$ may well be the dominant source of error, as Gigerenzer and Brighton (2009) argues that between-judge variance typically exceeds within-judge inconsistency, particularly when the decision is complex, meaning that there are many factors to weigh and the decision is not easily routinized (many "parameters" to estimate), and when the caseload per bureaucrat is small. Similarly, the distinction between whether the decision was erroneous or indeterminate from Fischman (2014) is of little importance to the actual citizen. Furthermore, systematic bias and arbitrariness also vary in remediability; where systematic bias can be perceived to be more unfair, it is more likely to be identified and challenged through legal or political channels, while arbitrariness from bureaucrat variance cannot.

6.3 Study 1: How much arbitrariness is there? Re-analysis of leniency-IV studies

Many studies have already studied arbitrariness without actually highlighting it. Studies using leniency instrumental variable designs all exploit arbitrariness through a common institutional feature: in many bureaucratic settings, cases are as-if randomly assigned to bureaucrats who differ systematically in their propensity to rule in favor or against a citizen (Chyn et al., 2025; Goldsmith-Pinkham et al., 2026). Assignment may be random by design (e.g. formal rotation

systems) or effectively random after conditioning on observable factors that are tied to the allocation process such as time periods, or particular courtrooms. The typical design defines $D_i \in \{0, 1\}$ as the binary decision for case i (e.g. incarceration, benefit denial, patent grant), Y_i as the downstream outcome, and $\mathbf{X}_i$ as a vector of observable case and timing characteristics. Each examiner j has a leniency measure Z_j, typically constructed as the examiner's average treatment rate across other cases they handle (a leave-out mean). These studies then use Z_j as an instrument for the endogenous decision D_i. The first stage regresses the decision on examiner leniency:

$$D_{ij} = \alpha + \beta\, Z_j + \mathbf{X}_i'\gamma + \varepsilon_{ij},$$

where β captures how much the probability of treatment shifts per unit increase in examiner leniency, conditional on case characteristics $\mathbf{X}_i$. The second stage then uses the predicted $\hat{D}_i$ from the first stage to estimate the causal effect of the decision on the outcome Y_i as a local average treatment effect (LATE) for the compliers, i.e., cases that could have received a different outcome. Therefore, there exists an entire body of research that, in a sense, has already measured arbitrariness in bureaucratic decisions, albeit without framing it as such.

Importantly, the identifying assumption is that examiner assignment is as good as random. This means δ_j is, by construction, orthogonal to case characteristics: it does not include private information or contextual judgment that a better τ would incorporate. The variation these designs capture is precisely the kind that cannot be rationalized as useful discretion.

Sample construction

I assembled the sample of leniency-IV studies from three sources. First, I started from the comprehensive list of 135 examiner and judge design studies compiled by Chyn et al. (2025), which covers published papers and working papers across criminal justice, finance, health, patents, disability, child welfare, and other domains. Second, I supplemented this with targeted searches using Elicit and LLM-assisted queries, which identified additional working papers and recent publications not yet included in the Chyn et al. list. Third, I conducted a systematic search on Web of Science using a broad set of terms covering examiner/judge leniency, stringency, fixed effects, random assignment, leave-out means, and related instrumental variable terminology (see Appendix D.16 for the full query). I screened all identified papers for whether they employ a standard leniency IV design, that is, quasi-random assignment of cases to examiners combined with an examiner-level leniency instrument as described above (typically a leave-one-out mean), and excluded studies using other identification strategies (e.g., difference-in-differences, natural experiments without an examiner instrument, or designs where examiner variation is the outcome rather than the instrument). This yielded 154 papers, of which I classified 127 as eligible, yielding 201 estimates across studies. The 27 ineligible papers were excluded because they did not use a standard leniency IV design despite involving quasi-random assignment to bureaucrats. Of the 127 eligible studies, 122 came from the original search pool: 106 originated from the Chyn et al. list, 5 were identified through LLM-assisted and Elicit searches, and 11 were found through the Web of Science search. Five were added later. My main analysis focuses on the 109 studies with binary treatments (e.g., detained vs. released, benefit approved vs. denied, patent granted vs.

rejected).

Some systems have been studied more than once, and I count each system once. When two papers study the same bureaucrats making the same decision, I keep the newest paper or the older paper if the data quality and reporting are significantly better. Appendix D.17 lists the removed studies: four papers at the screening stage and 15 estimates from 13 papers within the extracted sample. Papers that share a system but study a different decision or population are retained. For instance, the three patent-examination papers follow different applicant pools through the same patent office. The analysis sample is therefore 95 estimates from 88 papers.

I exclude the 25 studies with continuous treatments (e.g., sentence length, fine amount, benefit level) because their arbitrariness measures are denominated in incommensurable units, since months of incarceration cannot be compared to dollars of fines, whereas binary treatment ranges all live on the same zero-to-one probability scale.

As these studies are not meant to examine the extent and degree of arbitrariness, they do not always report the necessary ranges or first-stage statistics in a consistent way. In a few cases, I had to compute these statistics from raw data or replication files. However, this was rarely possible, as most of these studies use unavailable administrative data. Thus, I contacted authors for clarification or extracted data from digital scans of the figures when possible. My numbers therefore come from three sources: statistics reported directly in a paper's text or tables, values digitized from published figures, and statistics I computed from other reported quantities or from replication files. Appendix D.14 lists every study's numbers.

Estimands

From this type of design, I can gather two interesting types of estimands: (1) how often a new draw of bureaucrat would flip the decision and (2) how many cases could potentially go either way.

How many decisions flip? The first estimand measures the intensity of arbitrariness: if two eligible bureaucrats are drawn by chance, what is the chance that they would reach different decisions? Writing j and k for two independent draws from the pool J_i,

$$\text{GMD}_i = \mathbb{E}_{j,k} \left| D_{ij} - D_{ik} \right|, \tag{17}$$

which is the Gini mean difference of the potential decisions on case i. For binary decisions in the leniency studies, this is simply the chance that two draws of bureaucrats disagree, which is $2p_i(1 - p_i)$ when a share p_i of the pool would grant the case.

Here it does not matter which decision is right. Averaged over the caseload, $\mathbb{E}_i[\text{GMD}_i]$ is the chance that a randomly drawn citizen would have received a different outcome purely through the draw of who decided their case. Importantly, if we believe that every case has a correct decision, a correct τ, then one of the two bureaucrats is wrong, and the setting's expected error rate is therefore

$$\mathbb{E}_i[\text{Error rate}_i] \;\geq\; \frac{\mathbb{E}_i[\text{GMD}_i]}{2}, \tag{18}$$

so half the redraw share is a lower bound on the setting's error rate.

Sometimes bureaucratic decisions are really rare, e.g., taking children into foster care. For these decisions, it is relevant to ask a question from the perspective of the people who received them: of those who got the rarer decision, what share would a new draw of bureaucrat have decided the other way? In any disagreement between two draws, one draw gives the rarer decision and one the more common one, so this share is the disagreement chance from above divided by twice the share of cases receiving the rarer decision. This is the Gini coefficient of the bureaucrats' rates of the rarer decision. It also puts settings with very different base rates on the same scale: removal decisions that hit two percent of families can be read next to detention decisions that hit forty percent of defendants.

$$\Pr(D_{ik} \neq D_{ij} \mid D_{ij} = d^{\text{rare}}) = \frac{\mathbb{E}_i[\text{GMD}_i]}{2\,\min(\bar{p},\, 1-\bar{p})}, \tag{19}$$

where j is the actual, realized draw and k a new, random draw, d^{rare} denotes the less common of the two binary decisions across a study's cases, and $\bar{p} = \mathbb{E}_i[p_i]$ is the share of all cases granted, so $\min(\bar{p},\, 1-\bar{p})$ is the share of cases receiving the rarer decision.

When a decision is rare, even a small overall disagreement chance concentrates on the few who received it. For example, if bureaucrats disagree on one case in a hundred and two children in a hundred are removed, then, looking only at the removed children, one child out of four would have stayed with their family under a different draw.

How many decisions could flip? The second estimand measures the extent of arbitrariness: how many people could have received a different outcome? Thinking of the draw of bureaucrats as a lottery, this second estimand concerns how many people were in the lottery, whereas the first estimand concerns how many won.

This is known as the complier share in the IV literature (Angrist et al., 1996; Chyn et al., 2025; Goldsmith-Pinkham et al., 2026). Disregarding the always-takers, who always receive treatment, and the never-takers, who never do, how many people could potentially change treatment status depending on the draw of bureaucrat? Instead of averaging over the draws, I ask for the gap between the most lenient and the strictest bureaucrat. To compare between studies, however, I cannot rely on the most extreme-looking bureaucrats, since those are the noisiest points of the distribution and their gap grows mechanically with the size of the pool. I therefore report the nearest contrast I can estimate and validate: a bureaucrat in the strictest tenth of the pool against one in the most lenient tenth, the 10th percentile against the 90th:

$$\Delta_{10\text{–}90} = Q_{\bar{D}}(0.90) - Q_{\bar{D}}(0.10), \tag{20}$$

Here $\bar{D}_j = \mathbb{E}_i[D_{ij}]$ is bureaucrat j's treatment propensity over the case distribution. The function $Q_{\bar{D}}(q)$ returns the qth quantile of those propensities across bureaucrats. Thus $Q_{\bar{D}}(0.90)$ is the treatment propensity of the bureaucrat at the 90th percentile, $Q_{\bar{D}}(0.10)$ is the corresponding propensity at the 10th percentile, and $\Delta_{10\text{–}90}$ is the difference between them rather than the min–max range.

Every study uses the same identification strategy and measures the same thing, namely the

spread of its bureaucrats' decision rates. We can compare settings for both estimands, since they are shares of cases, so five percent means the same in a bail court as in a patent office.

The estimands are lower bounds Whether the range of examiner propensities can be read as a measure of arbitrariness relies on the monotonicity assumption: there are no defiers, meaning that there can be no case where a lenient examiner would deny but a stricter examiner would approve (Angrist et al., 1996).[7]

When we compare two examiners' approval rates, defier cases cancel against the ordinary complier ones, so the gap between the rates counts the ordinary disagreements minus the defier ones, while the share of cases the two would truly decide differently counts both together.

Both estimands are therefore lower bounds on true disagreement. For any pair of examiners, the share of cases they would decide differently is at least the gap between their approval rates, and it is exactly that gap under monotonicity (Fischman, 2014); the gap is the pair's complier share (Goldsmith-Pinkham et al., 2026). This applies both to the first and second estimand.

How much do the bounds understate? This has been measured by Sigstad (2026) in five settings with judicial panels. Here, several judges vote on the same case, so defier votes can be counted rather than assumed. I examine these counts at two levels, and these differ because one case holds many comparisons: with five judges on a case there are ten pairs of judges, and if one pair disagrees then the entire case is noted as being a disagreeing case, since it violates the strongest form of the monotonicity assumption. Counted by case, violations therefore look common. In 35 to 49 percent of the cases where the judges do not all agree, at least one pair disagrees against the judges' usual leniency ranking. But a single such pair then treats the whole case as a violation. If instead I count the disagreements, defiers are rarer. In the main São Paulo sample, 15.4% of disagreements go against the ranking. I can therefore do some light calculations and see that, out of every 100 disagreements, 15 defier ones cancel against 15 ordinary ones. A gap therefore shows about 70 of the 100. The true share of cases a pair would decide differently is thus the gap times about 1.4[8]. The estimands are lower bounds, but if examiners in this analysis resemble his judges, the true level is about forty percent larger than the numbers I report.

A further caveat for cross-study comparison is that the estimands depend on the composition of the case pool. In theory, studies could mechanically compress the range between bureaucrats, if it added cases that every examiner would approve (always-takers) or deny (never-takers), thereby increasing the denominator, while leaving the number of complier cases unchanged. The same study could thus define its sample more or less narrowly, e.g., screening cases to focus only on contested types, and would then show a larger complier share than studies that include all cases entering the system. Two things help with this worry. First, the question is system-level: what fraction of cases, as they actually enter the bureaucracy, are arbitrary? Second, the rare-side estimand is not affected by this, since dropping cases that would never

[7]Most studies test monotonicity themselves: 75 of my 109 binary-treatment studies split the sample on observable characteristics (defendant race, age, crime type, prior record) and check that a more lenient examiner raises the treatment probability in every subgroup; 69 of the 75 pass, and the remaining 34 assume monotonicity without a formal test. However, Sigstad (2026) shows these tests are weak and studies instead rely on average monotonicity (Frandsen et al., 2023).

[8]Sigstad (2026) does not do these calculations, but he presents his defier shares in Table 3.

receive the rare decision shrinks its numerator and denominator by the same factor and leaves it unchanged.

Estimation

As mentioned, very few of these studies have open replication data. From every study I collect different types of numbers: the spread of its bureaucrats' decision rates (standard deviations, stated percentiles, or values digitized and read off a published figure), the share of cases granted, the number of cases per bureaucrat, the number of bureaucrats, and the first-stage coefficient with its F statistic. These metrics are not available in every paper. Appendix Table 28 lists how often each input is available, where it comes from, and how it is checked.

There are three main challenges before I can actually get the estimands of the previous section. First, the estimands cannot be read off the papers: they have to be calculated from the spread of the bureaucrats' rates, so I need each study's true spread. Second, to compare different studies with different numbers of cases and bureaucrats, I need to remove the part of the visible spread that comes from noise rather than from the bureaucrats themselves. Third, studies build their leniency measures differently: some report raw rates, others report rates that are residualized, shrunken, or grouped, and some compute them on a different sample than the one they analyze.

The first challenge comes from not having the underlying data. The studies themselves are not concerned with the purposes of this paper, so they often do not report the statistics that I would like to compare, e.g., no study prints the gap between every pair of bureaucrats. However, if I assume that the bureaucrats' true rates follow a normal distribution, I can get the estimands and then validate using the few studies for which the full data are available. Both estimands can be expressed as multiples of the standard deviation, σ, of the normal distribution. The first estimand, the chance of disagreement, can be seen as the average gap between two randomly drawn bureaucrats, which is then $2\sigma/\sqrt{\pi} \approx 1.13\,\sigma$. The conditional share of equation (19) follows by dividing by twice the rare-side rate. The second estimand, the complier share or the gap between the 90th and the 10th percentile of the rates, is $2.56\,\sigma$.

I validate this using studies for which I have replication data or a full distribution from fine-grained digitized plots. The numbers are typically off by between one and five percent (see Appendix D.13). Therefore, for every study, I just need the true spread σ of its bureaucrats' rates.

The second challenge is getting this without estimation noise. A bureaucrat's measured rate that I want to compare is an estimate, and estimation adds some noise that could be taken for disagreement. For example, a bureaucrat who would grant 30 percent over a million cases will most likely not land precisely on 30 percent if they decided only a few hundred, simply because they were assigned some particular cases. Under as-if random assignment, the leniency measure is the bureaucrat's propensity measured with error, $Z_j = \bar{D}_j + \eta_j$, where η_j is the estimation noise, so the observed spread of the Z_j is wider than the true spread of the $\bar{D}_j$ (Goldsmith-Pinkham et al., 2026). I can correct this using the same correction used to measure dispersion in teacher quality (Kane and Staiger, 2008; Chetty et al., 2014). A rate estimated

from $\bar{n}$ cases has the sampling variance $\bar{p}(1-\bar{p})/\bar{n}$, and noise and true differences add in squares, so

$$\hat{\sigma}^2 = \mathrm{SD}_{\mathrm{obs}}^2 - \frac{\bar{p}\,(1-\bar{p})}{\bar{n}}, \tag{21}$$

where $\mathrm{SD}_{\mathrm{obs}}$ is the reported spread and $\hat{\sigma}$ estimates the true spread σ, the parameter I am actually interested in.[9]

The third challenge is that the reported spreads are not all the same object: some studies report raw leave-one-out rates, others residualize on controls such as court and year, and some report rates that are already shrunken, fitted from a model, or grouped (Goldsmith-Pinkham et al., 2026). I therefore recorded every study's instrument construction and calculated the spread accordingly (Appendix D.12). In practice, the types are comparable: raw and residualized instruments show no significant differences in their spreads (Appendix Table 29).

In total, I can get the numbers for 110 of the 118 estimates (101 of 109 papers). The eight remaining estimates, which use designs that do not report any bureaucrat-level spread, are documented in Appendix D.14.

To get confidence intervals I bootstrap using the bureaucrats' rates with the corrected spread and add the noise their caseloads imply (See Appendix D.12). Appendix D.14 furthermore lists every study's input values: the spreads the studies themselves report and the min-max and P10–P90 ranges of estimated rates; all of these raw values keep their noise component.

6.4 Results

How many decisions are arbitrary?

As an overall summary, across 95 estimates from 88 papers, the expected share of decisions that would flip had the case been reassigned to another randomly drawn bureaucrat is 5.8% at the median study (95% bootstrapped CI: 4.5% to 7.6%). At least 3.3% of cases would flip in 75% of all the settings in this analysis, as shown in Figure 8.

This shows how many errors the bureaucracies produce, purely due to arbitrary factors that decide the allocation of bureaucrats, if I assume that a case has a correct answer (Fischman, 2014). With this benign assumption, whenever two randomly drawn bureaucrats decide differently, at least one of them is wrong. So, because I cannot tell which of the two is correct, the median GMD of 5.8% suggests that about 3% of all decisions are errors.

Second, how many cases were in this lottery and could have been different? Comparing a lenient-decile with a strict-decile bureaucrat, I get a conservative count of the complier share: cases that could have received a different decision. Here, the median is 13.1%, and a quarter of settings exceed 29%.

Third, for the lucky and unlucky ones, i.e., concentrating on the people who received the rarer of the two decisions, the median share whose outcome a redraw would flip is 14.4%. This means that roughly one in seven of those removed, detained, or denied were treated so due to the assignment of a bureaucrat and not the facts of their case.

[9] See Appendix D.12 for practical details.

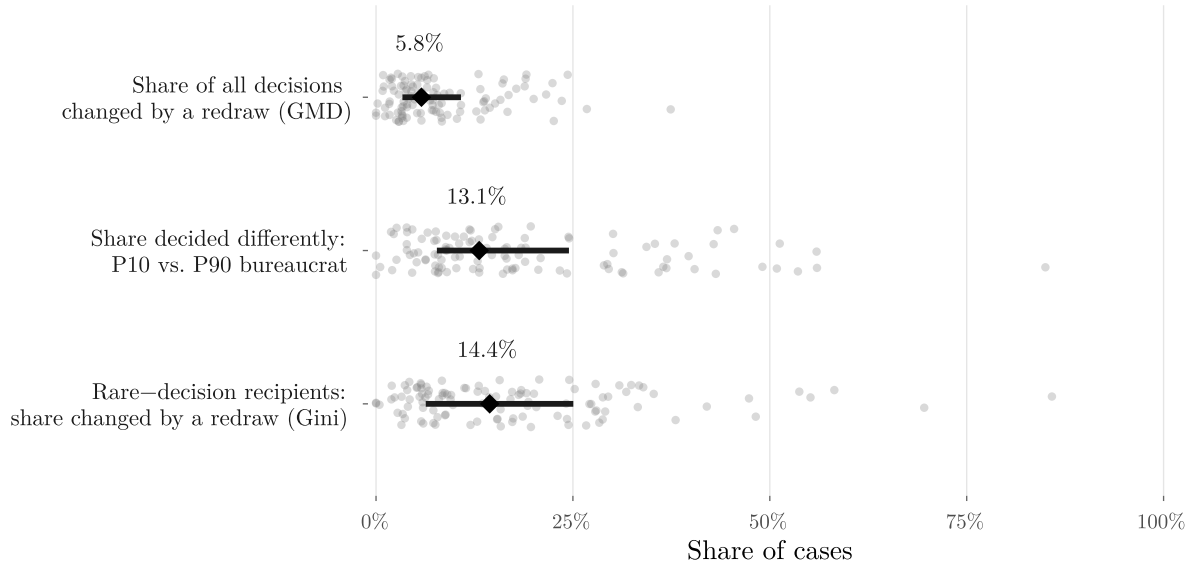


Figure 8: The three corrected estimands across the 95 estimates in the analysis sample (re-studied systems removed, subsection D.17). Each grey point is one estimate; the diamond marks the median and the black bar the interquartile range. All three numbers remove the spread that caseload noise alone would create and are lower bounds under the designs' monotonicity assumptions.

These numbers assume monotonicity, so each is a lower bound. They are also lower than the spreads the papers themselves report, shown in Figure 33 and Figure 34 in the appendix, because the reported spreads still contain the caseload estimation noise; in the median study, about a fifth of the visible, initially apparent spread is this noise. These measures are therefore the lowest reading of the numbers that is still reasonable.

Looking at the distribution of the redraw measure in Figure 9, most bureaucracies cluster at lower values but with a long tail to the right. For the modal study, the number of people who would have received a different outcome is 3 to 4% of the total number of cases in that setting; half of all studies fall inside the band from 3.3% to 12.9%. The mean of 8.3% is well above the median of 5.8% because of the tail. In the typical bureaucracy, a redraw thus changes about one decision in seventeen; in the most arbitrary setting, it changes one in three.

The tail is not noise; it consists of real settings with a high amount of arbitrariness, wherein the numbers are representative for it. In Figure 10 the confidence intervals are fairly short compared to the gaps between studies, so the differences reflect actual differences between settings. There is not a clear pattern in the tail. The three with the highest over 25% GMD are judicial foreclosure (Munroe and Wilse-Samson, 2013) and the anti-corruption court in the Philippines (Axbard, 2024) and just below 25% pretrial detention in Colombia (Diaz and Salas, 2022). No systematically higher level of arbitrariness appears in non-Western systems. Brazilian bankruptcy courts and Brazilian pretrial detention are in the upper middle (Araujo et al., 2023; Machado Ribeiro, 2019), pretrial detention in Argentina and business grants in Colombia sit on either side of the median (Di Tella and Schargrodsky, 2013; González-Uribe and Reyes, 2021), and the lowest estimate in the whole sample is non-Western as well. Also, the upper tail includes Western settings such as Michigan felony sentencing (Harding et al., 2019), land allotments in

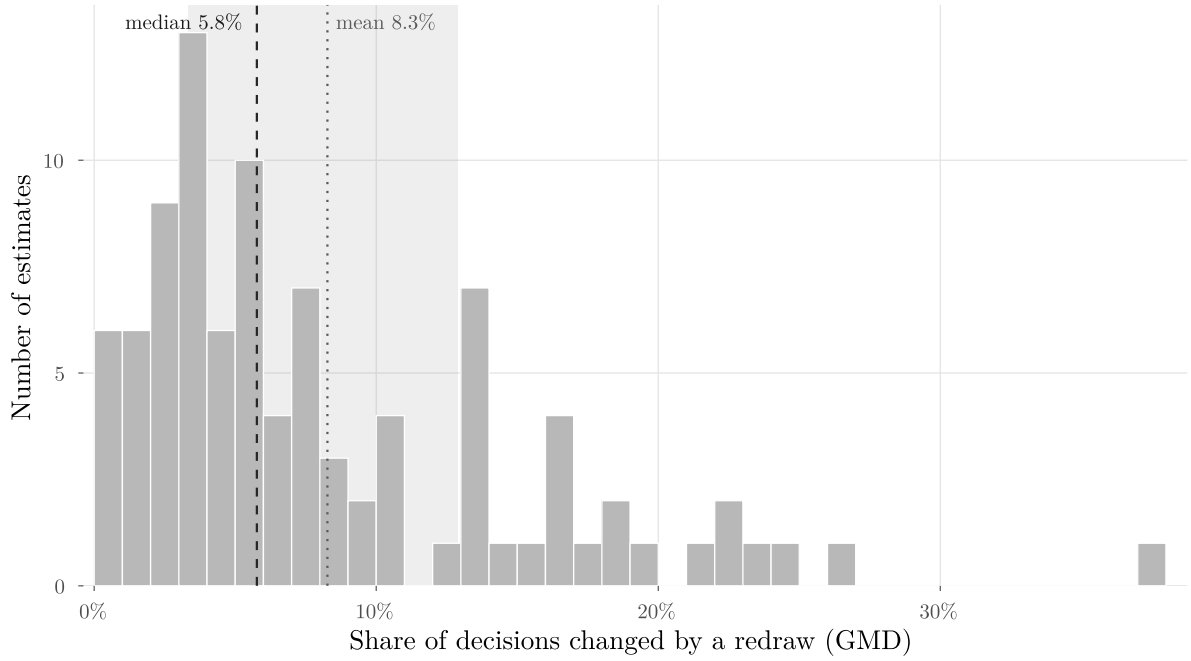


Figure 9: Distribution of the share of decisions that redrawing the bureaucrat would change (the corrected GMD), across the 95 estimates in the analysis sample. The shaded band is the interquartile range, 3.3% to 12.9%; the dashed line marks the median of 5.8% and the dotted line the mean of 8.3%.

India (Dippel et al., 2025), corporate bankruptcy in Portugal (Bonfim and Nogueira, 2024), and exam grading in California (Famulari, 2012), where two randomly drawn bureaucrats would decide differently in around one case out of five.

At the bottom of the distribution, there are many studies that are likely very protocol-driven. These have shares below 2%. They include, for example, radiologists reading chest X-rays (Chan et al., 2022), opioid prescribing in primary care (Eichmeyer and Zhang, 2023), and arrests at domestic violence calls (Amaral et al., 2023). There are, however, also studies such as Arteaga (2023) on incarceration in Colombia, Cortés (2025) on juvenile incarceration in Chile, and Dobbie et al. (2017) on bankruptcy judges: court settings with wide formal discretion where measured disagreement is nonetheless tiny.

There is not a single domain that dominates either end of the distribution; for instance, criminal justice appears across the whole range.

Another way of defining arbitrariness is not whether one's case was erroneous but whether it could have been: is there a bureaucrat who would have decided the case differently and might instead have received it? Therefore, I want to know the full extent of the lottery, or the full complier share as it is known in the leniency IV studies (Imbens and Angrist, 1994; Angrist et al., 1996). The median estimate is 13.1%, about one case in eight. In a quarter of the settings, this estimate exceeds 29%, and in the upper tail it exceeds half of the cases. In the bottom 10% of studies, 6% or fewer could be lucky or unlucky, depending on the setting. This is, however, a really conservative estimate both because of the monotonicity assumption and because I exclude the 20% most extreme bureaucrats in the tails. Had I looked at the difference between the P5 and P95 bureaucrats, the median would be 16.8%, and it would exceed 37.4% in the 25% most

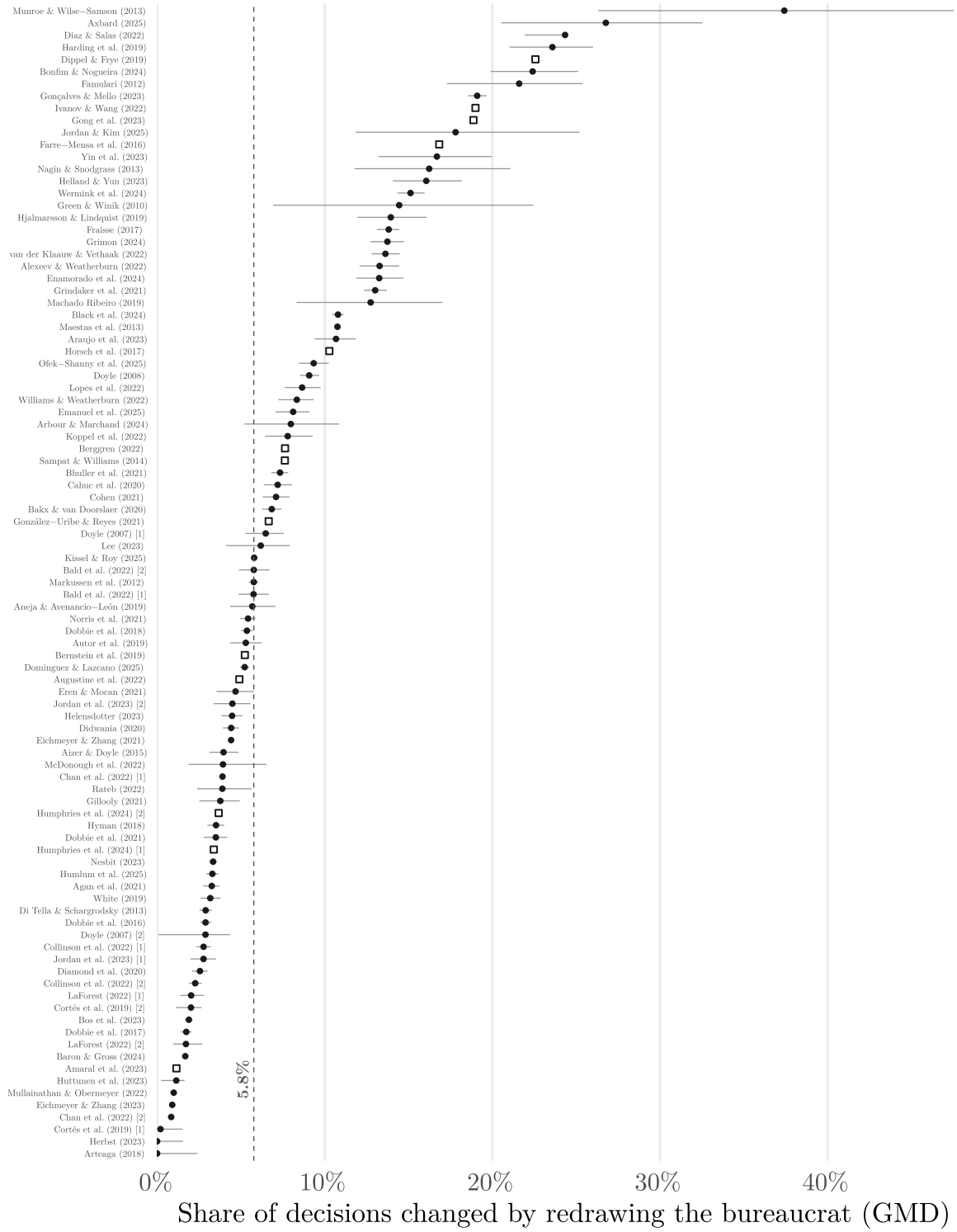


Figure 10: Share of decisions changed by redrawing the bureaucrat. Each point is the corrected GMD, i.e., the expected share of cases that two randomly drawn bureaucrats would decide differently, after subtracting the spread that caseload noise alone would create. Whiskers are 95% confidence intervals from bootstrapping; open squares mark estimates without such an interval. The dashed line marks the median of 5.8%.

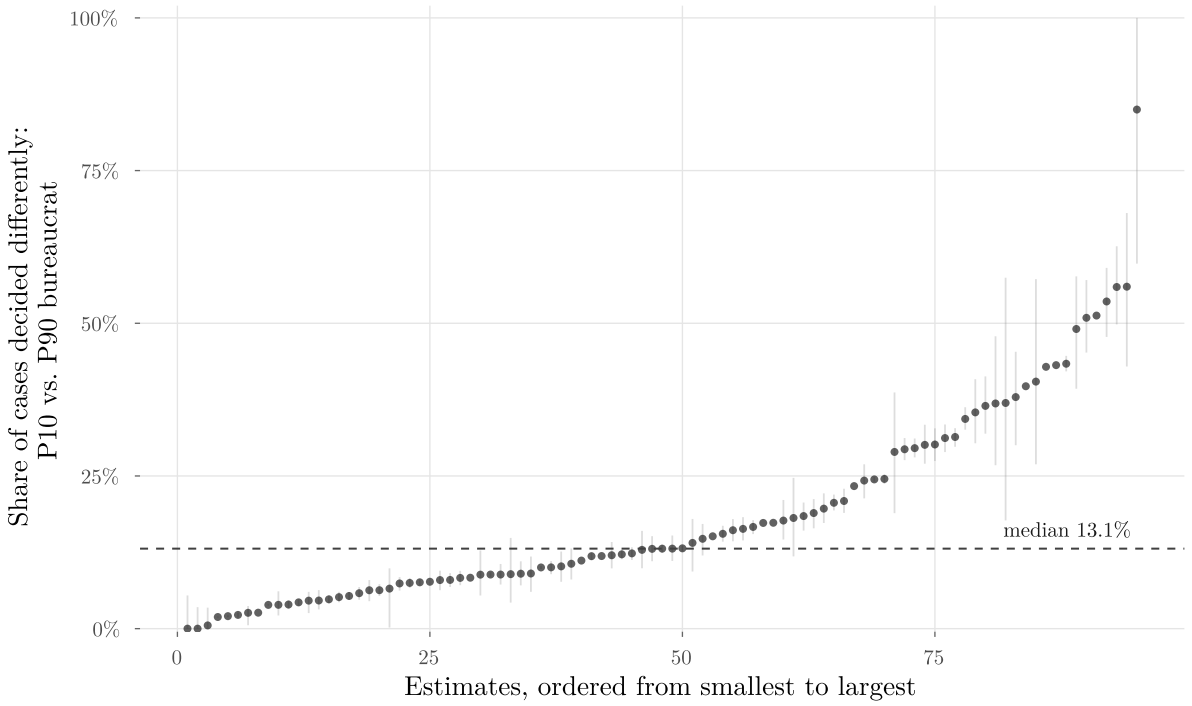


Figure 11: Share of cases a lenient-decile and a strict-decile bureaucrat would decide differently. Whiskers are 95% intervals; the dashed line marks the median of 13.1%.

arbitrary settings.

The ordering is the same as in Figure 10, since I estimated it from the same corrected spread. In Figure 33 I report the uncorrected spreads that are a bit wider and vary more across settings.

For the lucky and unlucky ones

In many bureaucratic decisions, one outcome is rarer than the other; children are more likely not to be removed than removed in foster care studies (Helénsdotter, 2026; Baron and Gross, 2025), defendants are more often released than detained (Dobbie et al., 2018), and disability appeals are more often allowed than denied (Black et al., 2024). To reflect this, I ask: conditional on receiving the rare outcome, what was the probability that the recipient could have received the other one? For the removed child, the detained defendant, and the denied applicant: were they removed, detained, and denied erroneously?

At the median study, 14.4% of rare-decision recipients would have received the other decision had the case been reassigned to another randomly drawn bureaucrat, and in a quarter of studies this share exceeds 27% (see Figure 32 for the full forest plot). This number is the Gini coefficient of the distribution.

For people receiving very rare decisions, there is a large amount of arbitrariness, as also shown in Figure 12. For the foster care children in Baron and Gross (2025), 2% were removed, and of those, about two out of five would not have been removed under another randomly drawn investigator. If there was a correct answer for each child to the question of whether they should have been removed, 20% were removed wrongly. Conversely, very few of the children who were

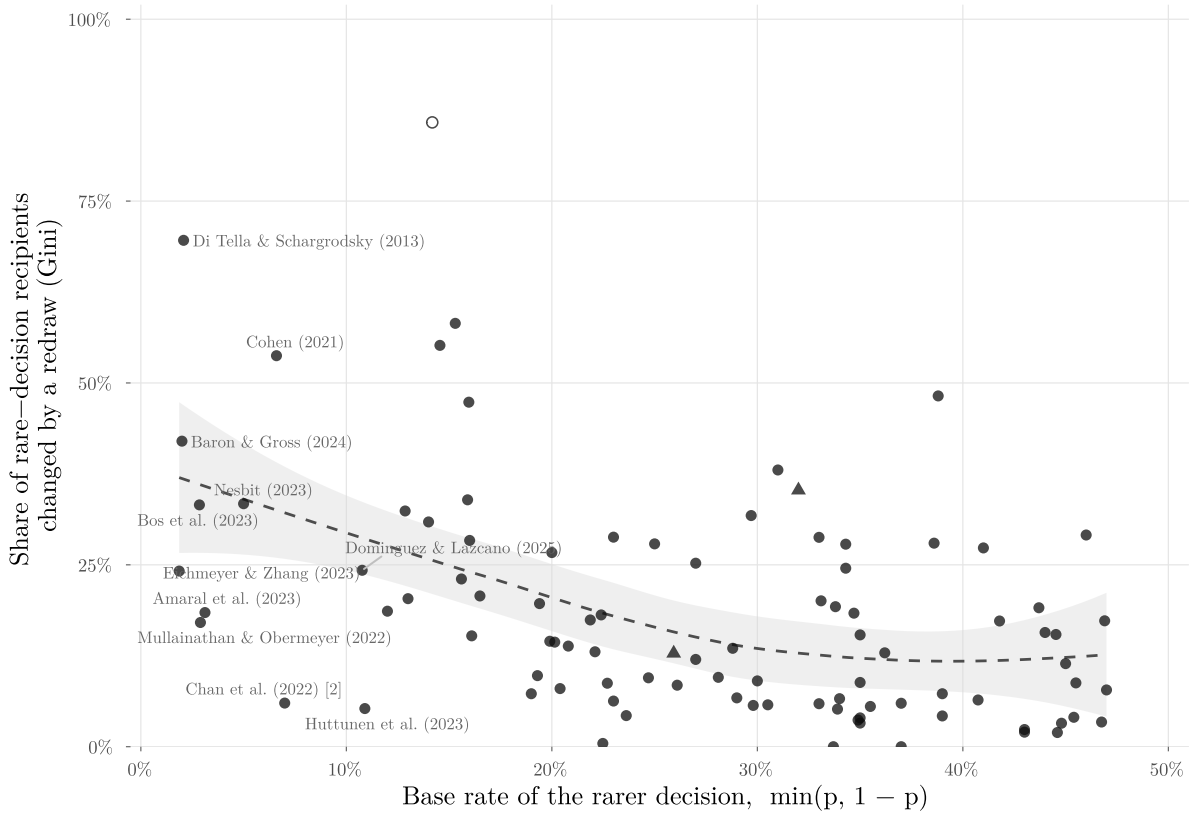


Figure 12: Of those who received the rarer decision, the share whose outcome a redraw of the bureaucrat would change (the Gini coefficient of the corrected decision-rate distribution), against the base rate of the rarer decision. The dashed curve is a loess fit with its 95% band; labeled studies have base rates below 12%; the open circle is a value capped at the feasibility bound of its base rate.

not removed should have been removed, at least as far as disagreement can reveal: under another draw, fewer than 1% of them would have been removed, so this same bound reveals an error rate of about 0.4%, roughly fifty times smaller than among the removed. The paper does not highlight this, but instead focusses on whether there was a difference between the compliers in their later incarceration rate.

The same type of inference can be made for other settings where the outcome is rare. For example, in Los Angeles, about half of the 7% of homeless individuals who get into a Housing First program owe the slot to the bureaucrat draw (Cohen, 2024). Not all settings where the outcome is rare have a large amount of arbitrariness, however. For instance, in a very routinized setting, the share is small: 6% for pneumonia diagnoses from chest X-rays (Chan et al., 2022).

At the extreme end, for a handful of settings the rare outcome is mostly due to the draw. In Jordan and Kim (2025), investigations of police misconduct complaints yield a supporting affidavit in about 15% of cases, and 58% of the obtained affidavits would not exist under another randomly drawn investigator. The paper is about police misconduct, but it also reveals errors by the very people, the investigators, who are supposed to prevent that misconduct in the first place. In Dutch criminal sentencing, 16% of cases end in short-term imprisonment, and 47% of those imprisoned would have avoided prison under another judge (Wermink et al., 2024); nearly

one in four of these imprisonments is an error. In five of the 95 settings, more than half of the rare decisions would be different under a redraw of bureaucrats.

Settings with very little disagreement can still leave the rare outcome largely to the draw. In Sweden, Bos et al. (2023) study conscription screening, where fewer than 3% of conscripts receive a mental health diagnosis and a redraw changes under 2% of all decisions, yet one third of those diagnosed would not be under another screening doctor.

There is thus a lot of luck in bureaucratic decisions. Denials show the same pattern at higher base rates: 31% of the 14% denied supported-housing applications in the Netherlands (Lopes et al., 2022) and 34% of the 16% denied disability appeals in the United States (Black et al., 2024) would flip under another bureaucrat. The same logic runs in reverse when the rare outcome is the good one, as with the Housing First slots above. In the Dutch unemployment insurance system, avoiding a job-search sanction is the rarer outcome at 27%, and one quarter of those who avoid it do so solely due to the draw (Klaauw and Vethaak, 2022).

What could explain arbitrariness? Discretion and preferences

Arbitrariness requires two things: (1) bureaucrats must have sufficient discretion to reach different decisions on the same case, and (2) they must differ in how they use that discretion.

There are many ways to design a bureaucracy, and deciding how much discretion legislators grant bureaucrats is a key parameter. Huber and Shipan (2002) argue that granting discretion is a deliberate institutional choice: legislatures can, e.g., write vague statutes when the policy domain is complex and when other oversight mechanisms such as judicial review or monitoring can constrain bureaucratic drift. Discretion is thus not a design flaw but a functional response to the impossibility of writing rules that anticipate every case (Lipsky, 1980; Zacka, 2017). The problem is that discretion also opens space for bureaucrats to diverge. As such, Bendtsen and Madsen (2025) find that longer and more detailed legislative texts are associated with fewer administrative errors, and Slough and Fariss (2021) find that providing more legal support to bureaucrats improves governance outcomes.

It is hard to distinguish among the leniency studies and find measures of discretion across them. There is some evidence that guidelines reduce the dispersion between judges (Yang, 2014; Anderson et al., 1999), the argument here being that the judges had less discretion in using them. The data therefore do not allow for a test of discretion.

Instead I can test whether bureaucrats have preference heterogeneity, so that they would differ when they have discretion, which is also the purpose of Study 2.

First, though, I can test whether observable characteristics of the settings themselves differ systematically. Although judges and other bureaucrats both decide cases according to rules, judges are typically more removed from politics, so I could expect them to differ more than bureaucrats who follow rules more directly through a chain of delegation.

Furthermore, I can distinguish between different types of decision-making. This is what Zacka (2017) does when he classifies three bureaucratic dispositions: enforcers, caregivers, and the indifferent. Whereas much of the literature follows Lipsky (1980) in arguing that bureaucrats' behaviour is determined by the institutional design of their working environment, Zacka instead

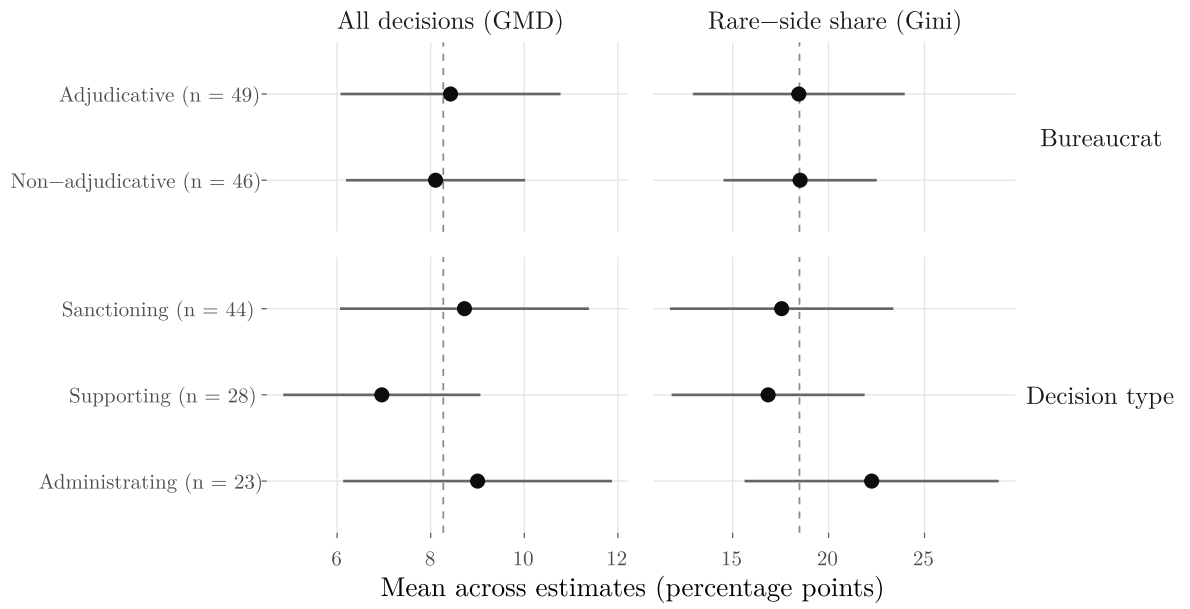


Figure 13: Mean corrected estimands by bureaucrat type and decision type, across the 95 estimates in the analysis sample. Points are unweighted means, in percentage points; whiskers are 95% intervals with standard errors clustered by paper; the dashed line marks the overall mean. Adjudicative bureaucrats are judges, magistrates, arbitrators, and tribunals. Decision-type definitions follow Table 24.

focuses on these individual dispositions and how they emerge from how bureaucrats cope with the moral weight of their work. Such dispositions shape how the same rules are applied differently to the same cases, producing systematic differences in outcomes across bureaucrats. Inspired by this, I can look at three corresponding tasks that reflect these dispositions: whether sanctioning, supporting, or administrative decisions are more arbitrary.

I find few robust patterns. In Figure 13, adjudicative and non-adjudicative bureaucrats are indistinguishable on both measures, and the three decision types overlap. The absence of strong covariate patterns can be seen as informative though. I cannot detect institutional configurations that reliably concentrate or eliminate arbitrariness although the statistical power is low.

Together, this highlights that it is hard to point out more arbitrary settings at face value. Rare outcomes, which one could argue are the ones most likely to drastically affect a citizen's life, are the most arbitrary, but the cross-study values are fragile, so this should be read with caution. This highlights, however, that arbitrariness is a part of bureaucracies' decision-making, since most settings have at least some amount of arbitrariness.

How much can selection move these numbers? The numbers above are basically summary statistics of many counterfactual settings. These settings, however, are not randomly chosen; they come from studies that use variation in bureaucrat assignment to estimate downstream effects. The assumptions behind the estimands bias each reported number down, which makes them lower bounds within their own settings. For the summary statistics, the medians and the IQRs, selection biases the numbers up in the other direction by shaping which settings enter the sample at all and favoring settings with disagreement. So, the sample could overstate how

arbitrary bureaucratic decisions are in general if there is a lot of selection.

There are primarily two ways in which this logic applies. (1) The design needs a strong first stage, so the weak-instrument convention, a first-stage F of at least 10, screens out settings where examiners nearly agree, and it does so at the design stage, before any paper exists to be written. (2) Settings where this type of design is feasible are systematically different from other bureaucratic settings (see Appendix D.9 for further notes and more tests).

For the first concern, this mimics the problems with publication bias and conditioning on significance: i.e., a study is more likely to be published if its estimates cross the threshold of having a p-value under 0.05, and it can do so only by being sufficiently powered or having incidentally large estimates due to noise. Therefore, summarizing a literature will overstate the average effect. In leniency studies, a study will instead not be written if the power in the first stage, essentially the F-statistic, is insufficient to generate variation in the second stage. The typical threshold for instruments is $F \geq 10$. I can therefore do some of the same tests and the same type of weighting that are common in testing the sensitivity of estimates to publication bias.

I can therefore ask: how much disagreement does a study need before it can exist? The design needs examiners who disagree enough to clear at least the usual strength requirement ($F \geq 10$), so each study has a value for the smallest dispersion it could have detected, given its own size. For the typical study, the value is a GMD of 1.6, so at the typical scale of these studies, the requirement of having enough variation to run a 2SLS model can hide only settings where examiners very nearly agree. The typical study finds a GMD 3.6 times higher than the minimum it could detect.

I can also test the numbers' sensitivity to this type of selection by asking what the sample would look like if settings below their lines were simply never written up. I follow the correction logic of Andrews and Kasy (2019); I assume that every setting with an F-statistic above different thresholds will always be written and reported, and studies under that threshold will only sometimes be written and reported. From the studies above the threshold, I fit a curve and test how many settings should have fallen below the threshold versus how many actually did. From this, I get weights that I can use to reweight the low-threshold sample and then see how sensitive the numbers are. In Figure 14, I test different thresholds. The thresholds move the median of the GMD estimand from 5.8 to 5.2 percent.

Some settings can never appear in the sample. If these settings have less arbitrariness, I cannot infer from these studies to the general problem of bureaucrats not making a correct decision, however that is defined. Settings with a single bureaucrat and systems without random assignment cannot be tested. Nor can I separate estimation noise in settings with very few bureaucrats.

The estimates are still lower bounds; thus, only the summary statistics of the distribution of studies can be biased upward. They may be biased upward by selection, while the monotonicity assumption that all bureaucrats can be ordered along one dimension biases each estimate downward when bureaucrats have heterogeneous preferences across clients. To assess whether there is substantial heterogeneity, and thus how conservative the lower bound may be, I turn to Study 2.

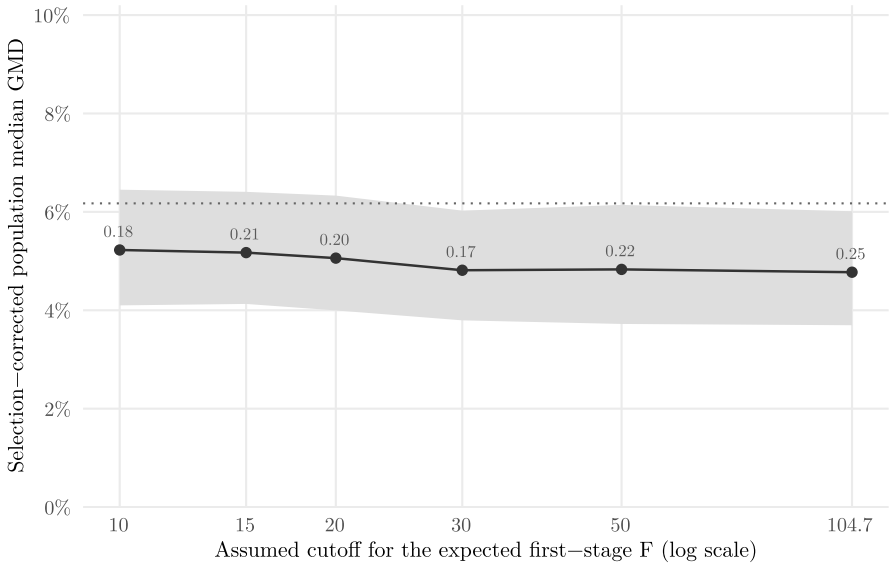


Figure 14: Selection-corrected median of the GMD at different thresholds, from the $F = 10$ convention to the strictest in the literature, 104.7 (Lee et al., 2022). The band is the bootstrap 95 percent interval, and the point labels are the estimated probability that a setting below the bar is written up anyway.

6.5 Study 2: Re-analysis and variance decomposition of conjoint experiments with bureaucrats

If there is room for different decisions and if bureaucrats have different preferences, then they will make different decisions, creating arbitrary decisions from the perspective of citizens. A central mechanism generating arbitrariness is preference heterogeneity. To test this mechanism, I would thus like to hold the case characteristics constant and see how much heterogeneity there is between bureaucrats vis-a-vis how much agreement there is. In other words, I would like to know: if cases are the same, would similar bureaucrats make the same decision, and just how much of the variance in decisions is due to bureaucrats weighing case attributes the same way, and how much is due to them weighing them differently?

This task is similar to that of Lauderdale et al. (2018), who asks how much of the variance in public opinion is due to (a) a single ideological dimension, (b) idiosyncratic but stable preferences, and (c) instability in responses. They measure the same respondents 1–2 months apart and then fit a model of their responses from a survey on public opinion. They model each respondent's position on a single ideological dimension, which predicts their responses on each issue to varying degrees (ideology), a respondent-by-issue random effect capturing stable preferences not explained by ideology (idiosyncrasy), and an error term (instability). They then compare how much each term explains of the total variance. To do the same and decompose variance in this fashion, I also need repeated measures of the same bureaucrats on the same questions to distinguish between simple response instability and stable differences between bureaucrats. The responses also need to be on the same scale to avoid measurement error.

Luckily, there is already a body of research that has sought to measure bureaucratic preferences with repeated measurements, namely studies using conjoint experiments. Unlike Lauderdale et al.

(2018), conjoint experiments do not have exact repeated measurements; however, they have the exact same attributes and the same type of question repeatedly measured across different cases. This lets me decompose variance in a similar way, albeit with some differences in the model specification and interpretation, so I can get a sense of how much of the variance in decisions is due to agreement between bureaucrats versus how much of the variance is due to heterogeneity between them.

First, however, I obtain a much simpler measure, the raw pairwise agreement. For each pair of respondents who evaluated the exact same comparison, I calculate the share that chose the same profile. Not all conjoint studies have sufficiently many cases where the respondents incidentally evaluated the exact same comparison. However, this is a model-free measure of how much bureaucrats agree when facing identical cases.

Second, the pairwise rate on its own conflates two distinct sources of bureaucrats making the same choice: either two bureaucrats pick the same profile because they share the same attribute preferences, or their idiosyncratic preferences happen to point the same way in that particular comparison. To address this, I decompose the variance in decisions using a random-slopes probit model that separates agreement in attribute weights from heterogeneity in those weights, similar to the one from Lauderdale et al. (2018). The decomposition is what tells these two sources apart, which is why the raw rate alone cannot say how much stable disagreement there is.

Essentially, I want to know (1) how similar bureaucrats are, which I call agreement, (2) how systematically different they are, which I call heterogeneity, and (3) how internally consistent they are, which I call consistency. In a sense, I am trying to estimate Equation 15, but the bureaucrat term can now vary with case characteristics rather than being a simple scalar of leniency.[10] I fit a random-slopes probit model:

$$y^*_{ir} = \mathbf{x}'_i\boldsymbol{\beta} + \mathbf{x}'_i\mathbf{u}_r + \varepsilon_{ir}$$

where y^*_{ir} is the latent utility that respondent r assigns to one profile over the other in comparison i, and $\mathbf{x}_i$ is the difference in case attributes between the two profiles. The first term in the model ($\mathbf{x}'_i\boldsymbol{\beta}$) is the agreement component: the population-average effect of each attribute on the decision. These coefficients correspond perfectly to the average marginal component effects (AMCEs) that are the standard estimand in conjoint analysis (Hainmueller et al., 2014). When this term is large in terms of variance relative to the others, bureaucrats respond to case characteristics in the same way. The second term ($\mathbf{x}'_i\mathbf{u}_r$, where $\mathbf{u}_r \sim N(\mathbf{0}, \text{diag}(\boldsymbol{\sigma}^2_u))$) is heterogeneity in bureaucrat preferences: respondent-specific deviations from the shared attribute weights. That is how they each systematically deviate from the AMCEs. This term therefore captures systematic differences in how individual bureaucrats weigh case characteristics that are stable across the comparisons they evaluate. Because each bureaucrat makes a choice in every comparison, there is no term for overall leniency; here, I only test the heterogeneity that would contribute additionally to the arbitrariness of each setting in Study 1. The third term ($\varepsilon_{ir} \sim N(0, 1)$) captures instability: variation in responses that is not attributable to either shared or individual attribute weights. This includes simple measurement error, but also inattention and within-respondent inconsistency. The observed choice is $Y_{ir} = \mathbf{1}[y^*_{ir} > 0]$.

[10] Using the language of Kahneman et al. (2021), heterogeneity corresponds to "stable-pattern noise."

I am not interested in the sign or magnitude of these terms for any individual respondent but rather in how much variance each term generates across respondents. Because the three sources of variation are uncorrelated, the total variance in the latent utility splits into three pieces that can be added together, so I can also get the percentage of the choice variance that each term explains. The terms I am interested in are therefore:

$$\underbrace{\text{Var}(y^*_{ir})}_{\text{total}} = \underbrace{\text{Var}(\mathbf{x}_i'\boldsymbol{\beta})}_{\text{agreement}} + \underbrace{\text{Var}(\mathbf{x}_i'\mathbf{u}_r)}_{\text{heterogeneity}} + \underbrace{1}_{\text{instability}},$$

and the share of each source is its piece divided by the total. The instability piece equals one because the probit scale fixes the error variance, which gives the three pieces a common yardstick (following Lauderdale et al., 2018). The derivation of the terms is reported in Appendix D.2. I fit the model in Stan (Carpenter et al., 2017).

While my approach mimics Lauderdale et al. (2018), due to the data structure my approach differs from theirs on all three parameters. First, in Lauderdale et al. (2018), the ideology component is a latent factor model: a single ideological dimension that loads differently on each issue. The agreement component replaces this with fixed attribute effects $\boldsymbol{\beta}$, so I do not impose a single latent dimension on bureaucrats' preferences, like conservatism versus liberalism, but instead let each case attribute have its own shared weight so it can correspond to the AMCEs. There is a near-perfect correlation between OLS-estimated AMCEs and my estimates when converting the probit coefficients to marginal effects (mean $r = 0.998$; see subsection D.4).

Second, in Lauderdale et al. (2018), the idiosyncrasy term is a respondent-by-issue random effect, i.e., a parameter for each respondent on each survey question. My heterogeneity term instead gives each respondent a random slope on each attribute ($\mathbf{u}_r$), meaning that individual deviations are structured by case characteristics rather than being question-specific. This is necessary because conjoint tasks are not exact repeated items. Each comparison involves a unique combination of attribute levels, so there is no fixed "issue" to which a random effect can be attached.

Third, Lauderdale et al. (2018) identifies instability from exact repeated measurement: the same respondent answers the same question across two survey waves. I lack such test-retest data as each conjoint comparison is unique, so my instability term is a residual that absorbs all variation not explained by shared or individual attribute weights, including any preference structure not captured by the additive model (e.g. correlational effects between attributes).

This differs from other methods for detecting heterogeneity in conjoints. For example, Robinson and Duch (2024) uses tree-based classifiers for specific covariates to explain heterogeneity, and Zhirkov (2022) does so by estimating an individual regression model per respondent when respondents complete 15 or more rounds of discrete choices. I am neither interested in specific covariates, nor do I have studies where respondents complete 15 rounds of comparisons. This method is computationally heavier; however, it allows me to measure the degree of heterogeneity directly without assuming it should correlate with anything.

Sample construction

I assembled a sample of conjoint studies through a systematic search on Web of Science combined with targeted searches. The Web of Science query targeted studies stating that they used conjoint or discrete-choice experimental designs with public employees, bureaucrats, or civil servants as respondents, returning 124 results. I supplemented these with targeted searches through Google Scholar and citation tracing and screened them on three inclusion criteria. The study had to (1) use a conjoint or forced-choice experimental design, (2) have bureaucrats or public employees as respondents (not citizens), and (3) have an experimental task that mimics a real administrative or case-level decision (not personal policy preferences, workplace attitudes, or abstract evaluations). After I applied my inclusion criteria, 26 of the 124 papers were eligible. The most common reason for exclusion was that the conjoint surveyed citizens rather than bureaucrats; other reasons included measuring workplace preferences or public service motivation rather than case decisions.

Where the studies did not have open data, I contacted the authors of all 26 eligible studies requesting individual-level replication data. Of these, I obtained data from 13 studies. The remaining 13 either did not respond to my request or could not share data due to confidentiality restrictions. 11 of these studies had shared comparisons between respondents, allowing me to compute pairwise agreement rates. In Jankowski et al. (2020) and Dietrich et al. (2026), there were no shared comparisons, so I could not compute agreement but could still include them in the variance decomposition analysis. For the variance decomposition model I had 17 experiment-level estimates, since some studies contain multiple experiments. After applying Bayesian convergence criteria ($\hat{R} < 1.05$, effective sample size > 100, no divergent transitions), 13 estimates from 10 unique studies remained for the main analysis. Three studies failed to converge because they had too many conjoint attributes relative to the number of respondents and iterations, making the random-slopes model insufficiently identified (Hollibaugh et al., 2020; Jankowski et al., 2020; Boittin et al., 2024).

Results

First, I look at the simple shared comparison. Across the 10 studies, the mean pairwise agreement rate is 61.4%, ranging from 54.4% to 74.1% (Figure 15). With random choices, we would expect 50% agreement in a binary forced-choice design, so agreement is consistently but not much above chance. When two bureaucrats are presented with the exact same pair of cases, they choose the same one roughly three out of five times. This leaves substantial room for disagreement even when case characteristics are held constant.

The study with both the highest pairwise agreement and the highest agreement share is Liu (2024), a study of partisan bias among city managers. While one might expect partisan considerations to divide respondents, the disagreements are not concentrated on the partisan attribute: among the pairs of bureaucrats who disagree, the two cases differ on partisanship less often than on any other attribute (see subsection D.5). That this study ranks highest likely reflects the simplicity of the design, with only four attributes, three of which (job creation, resource allocation, and collaborative trust) have a clear normative ordering.

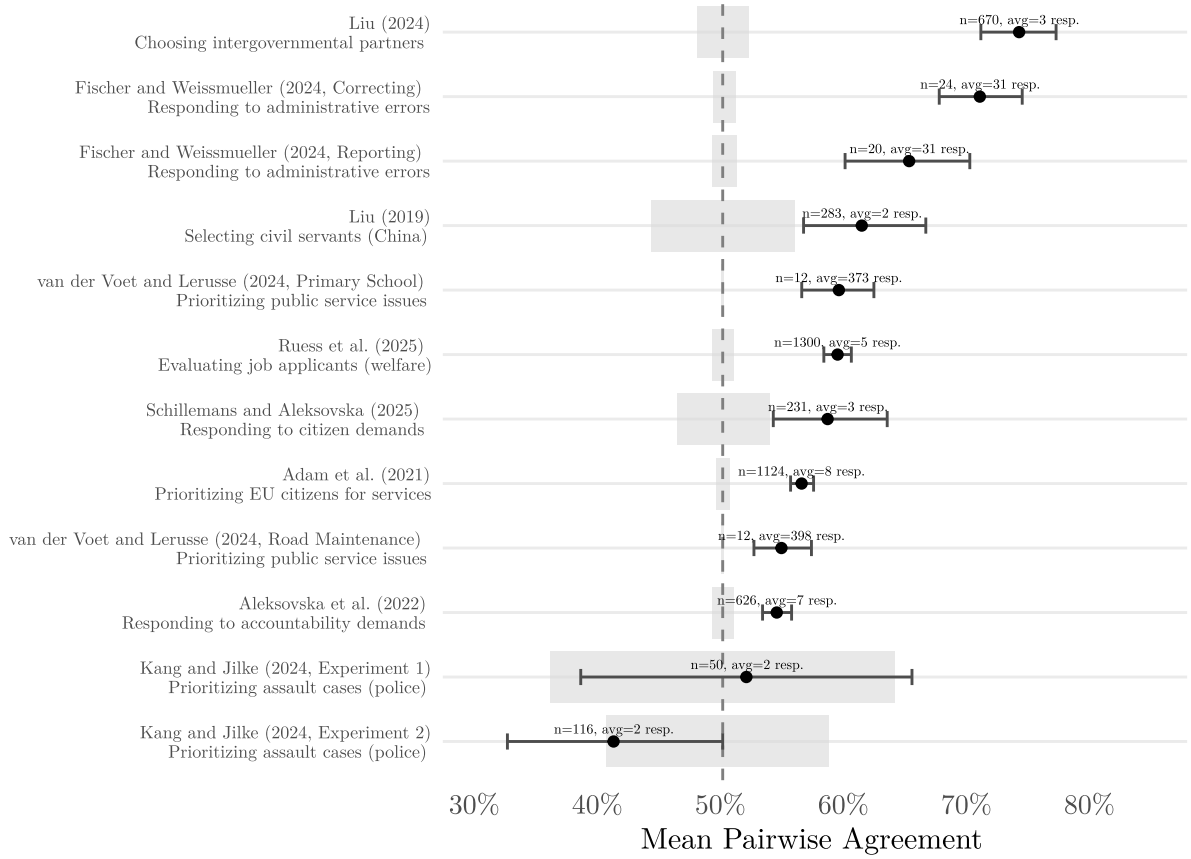


Figure 15: Pairwise agreement across conjoint studies. Black points show observed mean pairwise agreement with 95% bootstrap confidence intervals. Grey bands show the 95% simulated interval under random choice ($p = 0.5$), given each study's number of shared comparisons and average respondents per comparison. n = number of shared comparisons; avg = mean respondents per comparison. The Kang and Jilke experiments are excluded from the main table because each shared comparison has only two respondents.

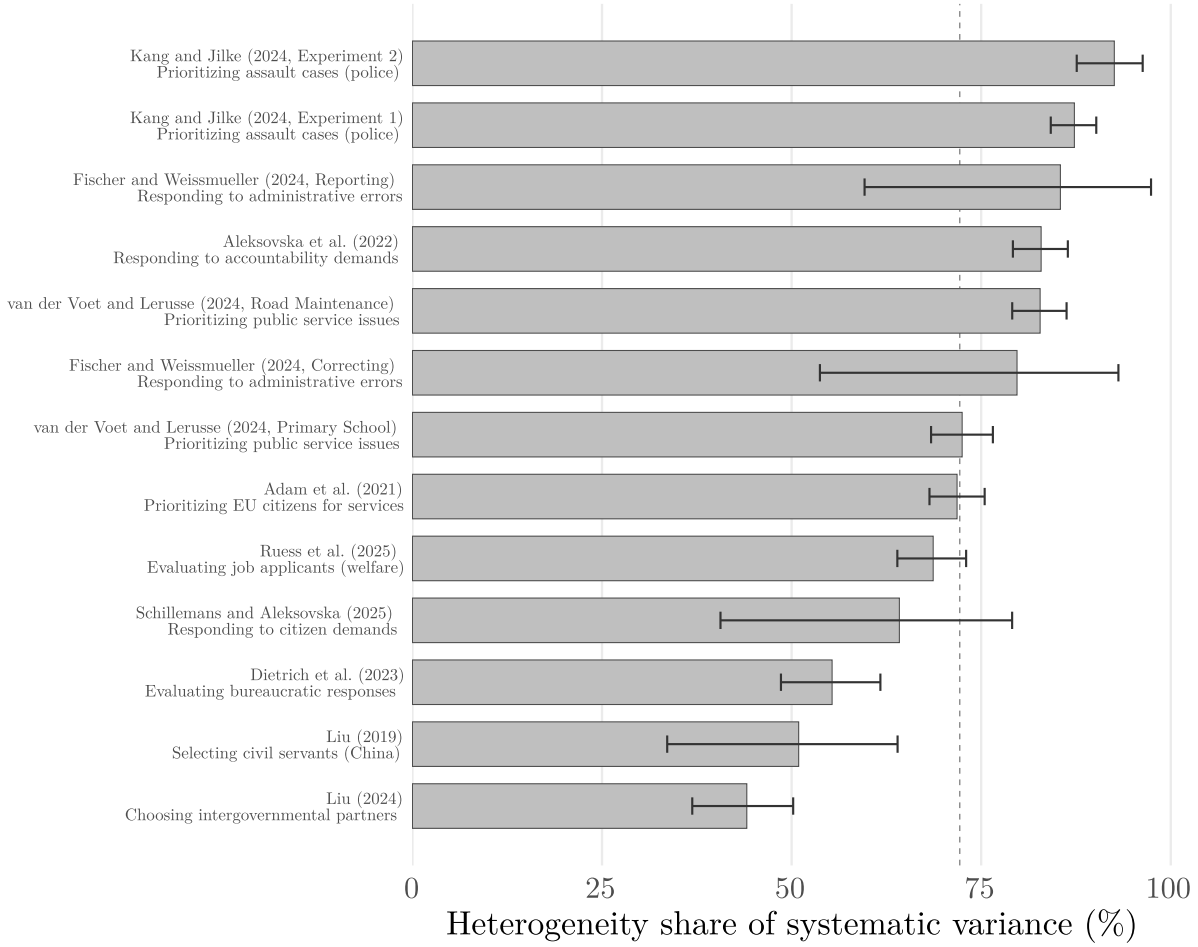


Figure 16: Between-respondent heterogeneity in conjoint responses, by estimate, with 95% posterior credible intervals. Each point is the share of systematic variance, with instability excluded, that comes from stable between-respondent differences in how attributes are weighted, that is, heterogeneity divided by agreement plus heterogeneity (σ_u relative to β plus σ_u). Intervals are the 2.5th and 97.5th posterior percentiles from the corrected random-slopes probit decomposition, in which agreement equals $\beta' \mathrm{Cov}(X)\beta$ and heterogeneity equals $\sum_k \sigma^2_{u,k}\, \mathbb{E}[X^2_k]$. The dashed line marks the unweighted cross-estimate mean. The corresponding shares of total response variance, including the instability (noise) share, are in Figure 45.

Two bureaucrats may pick the same choice because they share the same preferences in attributes, or because their idiosyncratic preferences just happened to point in the same direction. To separate those, I do the variance decomposition. Figure 17 plots the full decomposition of the total response variance for each experiment. The median shares are 17.1% agreement, 43.8% heterogeneity, and 29.0% instability, with a 95% credible interval on each bar. Figure 16 then compares the two systematic pieces directly, rescaling agreement and heterogeneity to sum to one with instability set aside. On this scale the median agreement share across the 13 experiment-level estimates is 27.5%, ranging from 7.4% to 55.9%. In 12 of 13 experiments heterogeneity exceeds agreement, often by a wide margin; the median heterogeneity share is 72.5%. Stable differences between respondents in how they weigh case attributes account for the majority of the variance that the attributes explain.

Across studies, agreement is highest when the decision task has attributes with a clear

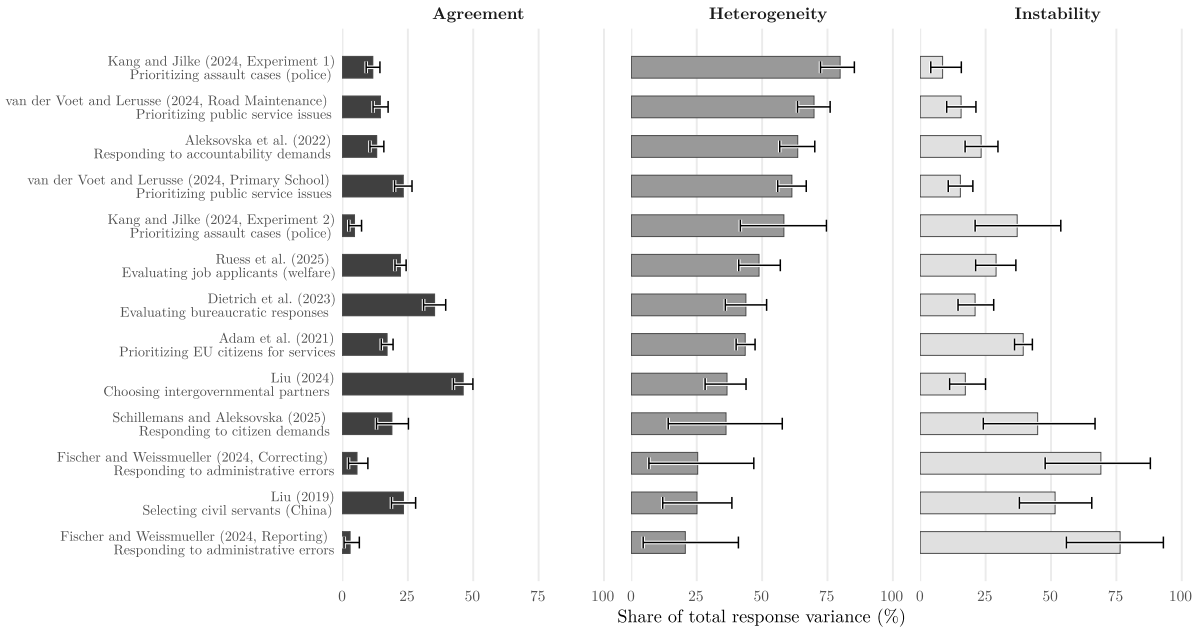


Figure 17: Shares of total response variance, by component, with 95% posterior credible intervals. Each bar is the posterior-mean share of total response variance for one estimate; whiskers are the 2.5th and 97.5th posterior percentiles. The three panels share the same scale, so bar lengths compare across panels. Agreement equals $\beta' \mathrm{Cov}(X)\beta$ and heterogeneity equals $\sum_k \sigma^2_{u,k}\, \mathbb{E}[X_k^2]$, each divided by the total variance, and instability is the fixed unit error variance divided by the total (subsection D.2); the three shares sum to one for each estimate. Estimates are ordered by their heterogeneity share.

normative ordering, and lowest when the task more apparently involves genuine value conflicts. The three studies with the highest agreement shares, Liu (2024) (55.9%), Liu (2019) (49.1%), and Dietrich et al. (2026) (44.7%), all present bureaucrats with cases where most attributes point toward an objectively "better" option: more jobs, higher financial assets, greater need, or better collaborative trust. Even among these, agreement exceeds heterogeneity only in Liu (2024); in the other two, heterogeneity remains the larger share. In contrast, the studies with the lowest agreement shares involve decisions where reasonable bureaucrats can disagree on principle. Kang and Jilke (2024) (7.4% and 12.7%) presents police investigators with sexual assault cases requiring subjective credibility judgments. Aleksovska et al. (2022) (17.1%) concerns transparency decisions where public accountability conflicts with institutional reputation. Fischer and Weißmüller (2024) (14.6% and 20.3%) asks bureaucrats whether to report or correct errors made by colleagues, a trade-off between accountability and institutional loyalty with no clear right answer.

These results mean that when bureaucrats evaluate the same case profiles, the largest single source of what drives their choices is not a common reaction to case characteristics but rather individual differences in how they weigh them. Two bureaucrats may both care about the same attributes, but weight them differently and thus produce different decisions for the same case. This suggests that there is significantly more arbitrariness than the lower bounds in Study 1 suggest. For the general theory of discretion, this is consistent with the leniency-IV findings: bureaucratic outcomes depend substantially on which bureaucrat you draw.

An important caveat is that conjoint experiments are designed to elicit trade-offs. Researchers select attributes where the ordering is not self-evident; nobody designs a conjoint where everyone obviously agrees. This means the studies may focus on dimensions where they suspect heterogeneity relative to real case characteristics. If the attributes that matter most in practice are ones where bureaucrats largely agree, then the conjoint-based heterogeneity estimates overstate real-world arbitrariness. Conversely, if real cases involve the same kinds of difficult trade-offs that conjoints present, the estimates may be realistic at least on average. With these results in mind, however, we can speculate on how much we should actually care about average effects if they explain so little of the actual choices that bureaucrats make.

6.6 Discussion and conclusion

Bureaucrats are important, and whoever gets to decide a citizen's outcome has a lot of say in what that outcome becomes. Across 95 binary-treatment estimates from 88 papers spanning criminal justice, disability determination, immigration, bankruptcy, and child welfare, at least 5.8% of cases would be decided differently in the median setting. Provided that there is a correct decision, this means that at least around 3% of cases in the median setting are decided erroneously, and around one in seven is at risk of being decided erroneously.

It is a lower bound because every case in which a more lenient bureaucrat would have denied an application that a stricter bureaucrat would have approved cancels out, and I cannot estimate these cases. To get a sense of these numbers, I hold leniency constant and look only at variation between bureaucrats while holding all case characteristics constant using 13 conjoint experiments. In these, facing identical case profiles, bureaucrats agree on only about three of five paired comparisons (a mean of 61%, a median of 59%), and of the choice variation that case attributes explain, the larger part reflects stable heterogeneity between bureaucrats rather than agreement, with a median heterogeneity share near 70% against 30% for agreement.

Whether the numbers should be seen as large or small depends on the individual context, but overall this indicates that arbitrariness in bureaucratic decisions is widespread.

What I cannot study is why bureaucrats differ. My estimates reveal one systematic difference between bureaucrats in overall leniency, but I cannot study the deeper sources of that divergence. Partisan motivation, moral orientation (what Zacka (2017) calls enforcers, caregivers, and the indifferent), differences in risk tolerance, or accumulated experience could all play a part. The variance in arbitrariness cannot be explained by measurable traits, but that might actually be suggestive. If arbitrariness could be reduced to a handful of characteristics, it could probably be addressed through targeted selection or training; the failure to do so suggests that it might simply be a feature of discretion in implementation.

I argue that to get a less error-prone, fairer bureaucracy, we need to look at factors other than average biases. A bureaucracy can be unbiased on average and still be arbitrary if its bureaucrats disagree about cases, so reducing the mean error and reducing the dispersion of decisions are distinct problems that call for distinct remedies. I thus make the simple argument that, conditional on the quality of a decision, predictability is what remains to be desired. My

two designs narrow what this variance can be. Random assignment in the leniency studies makes the aggregate composition of case characteristics equal, and the conjoint experiments hold case characteristics fixed by design, so in neither setting can the disagreement reflect bureaucrats knowing different things about the case.

How bureaucratic discretion can be exercised responsibly has, however, been a central debate in the study of bureaucracy, most notably between Friedrich (1940) and Finer (1941). On the one side, Friedrich (1940) argued that democratic control cannot rest on a simple notion of a unified "will of the people"; because public preferences are heterogeneous and administrative problems are technically complex, responsible administration must be elicited through responsiveness to both public sentiment and professional standards. On the other, Finer (1941) responded that the difficulty of identifying a unified popular will does not license replacing democratic control with bureaucratic professionalism. For Finer (1941), public servants have expertise in what is technically feasible, but they remain accountable to elected institutions, which rightly have the authority to define public purposes of the administration. Even granting that the target is underdetermined, it is doubtful that Friedrich himself would have accepted disparities of the size I document. His ideal still demanded responsiveness to shared professional standards, which implies some convergence.

The particular cost of arbitrariness is that citizens cannot see it for what it is. When the same application can be approved or denied depending on who reads it, citizens cannot learn how the system works, which exacerbates the administrative burdens they already carry (Herd and Moynihan, 2018).

What should we do to remedy arbitrariness? An intervention that lowers arbitrariness without changing how case characteristics map to correct decision outcomes (what I call τ in the equation) is unambiguously beneficial, whereas one that also reshapes τ faces a trade-off. Some studies have looked at encouraging more professionalism; e.g., Ho (2017) studied peer review among food-safety inspectors in an RCT and found that the dispersion in their scores dropped. Because this aligns how bureaucrats read the same facts rather than replacing their judgment with a fixed rule, it compresses $\mathrm{Var}(\delta_j)$ without obviously distorting τ. Alternatively, US studies examine the introduction of guidelines (Anderson et al., 1999) or their removal (Yang, 2014), which also influence disagreement between bureaucrats, in this case judges.

In the near future, reducing arbitrariness is possibly one of the great opportunities for using algorithms and AI. There are already studies on the effectiveness of using algorithms, e.g., Kleinberg et al. (2018), whereas the effects of using AI and LLMs are less predictable and should thus be studied further.









Chapter 7

Noise in Administrative Decision-Making: Public Administration's Unseen Error and AI's Role in Enhancing Personnel Selection Consistency

Co-authored with Mogens Jin Pedersen and Florian Keppeler

Abstract. Unwanted variability in judgments—or 'noise'—undermines administrative decision-making. Although as detrimental to equality as systematic bias, noise remains largely unaddressed in Public Administration scholarship. This article introduces the concept of noise to the field, distinguishing it from bias and highlighting its consequential (mis)alignment with core public values. Overlooking noise risks misattributing judgmental errors to bias, obscuring the role that random inconsistency can often play in driving unequal outcomes. We then empirically investigate how artificial intelligence (AI) can mitigate noise within public personnel selection, evaluating two human-AI collaboration modes: 'ensembling' (aggregating human and AI judgments) and 'specialization' (human judgments informed by AI advice). Our findings demonstrate both modes reduce noise. While ensembling's efficacy increases mechanically with the AI's weight, specialization achieves a behavioral noise reduction equivalent to an ensembling approach with a 13.2% AI weight (95% CI: [10.3%, 16.4%]). This study initiates a crucial research agenda on noise in the public sector.

Key words: noise; administrative decision-making; artificial intelligence; human-AI collaboration; personnel selection

Evidence for Practice

- To uphold the public value of equality, public managers should identify, measure, and manage 'noise'—unwanted variability in professional judgments—as a critical source of decisional error, alongside systematic bias.
- Implementing human-AI collaboration offers a practical pathway to enhance decisional consistency. This research demonstrates that two specific modes—'ensembling' (aggregating human and AI judgments) and 'specialization' (human judgments informed by AI advice)—are effective strategies for mitigating noise in high-stakes tasks such as personnel selection.
- Selecting a human-AI collaboration mode involves a strategic trade-off: While ensembling may offer greater noise reduction, specialization provides a consistency gain that prioritizes human decision-making authority.

7.1 Introduction

Consider the case of Ms. Albright, a public servant who adjudicates disability claims. On a sunny Monday, well-rested, she approves an application. Later that week, fatigued on a rainy Thursday, she denies a functionally identical case. Meanwhile, her colleague, Mr. Davies, consistently applies stricter standards to parallel applications. Such within-rater and between-rater variabilities fundamentally contradict the administrative principle of equality—"equal treatment of equal cases" (Raaphorst, 2021; Raaphorst et al., 2026)[11]—rendering eligibility contingent upon factors irrelevant to the merits of the case.

Such divergent outcomes, where judgments for similar cases vary within and between decision-makers, are instances of what has been termed 'unwanted variability' (Brighton and Gigerenzer, 2015) or 'noise' (Kahneman et al., 2016; Kahneman et al., 2021). Noise, in this context, refers to unwanted variability in professional judgments and predictions concerning cases that are, in principle, identical. As Kahneman et al. (2021) assert, "*wherever there is judgment, there is noise—and more of it than you think*" (p. 12). This variability, encompassing both within-rater inconsistency and between-rater differences across decision-makers, constitutes a significant, yet largely neglected, challenge in Public Administration (PA).

Much PA discourse has centered on mitigating systematic *bias* in administrative decision-making—a systematic error that, on average, skews decisions in a specific direction, often uncodified by or contrary to established policy or professional norms. Such distortions are therefore illegitimate and normatively unwanted, and are frequently attributed to the cognitive biases—attitudes, beliefs, and perceptions—of public employees. Yet, the random scatter of noise equally imperils the ideal of equal treatment. Drawing on insights from cognate disciplines (e.g., Brighton and Gigerenzer, 2015; Kahneman et al., 2021), we emphasize that noise contributes substantially to decisional inaccuracy, thereby eroding equitable service delivery and representing arbitrary (random) inconsistency from the perspective of the citizen.

Although decisions involve nuanced factors beyond mere judgment or prediction (Athey, 2017), professional judgment and prediction remain integral components of discretion and decision-making by public employees. Despite this centrality, PA scholarship has long problematized such variability yet without directly conceptualizing it as a distinct source of error. Indeed, a recent research synthesis indicates that noise remains a notable empirical blind spot in the field (Cantarelli et al., 2023), having not yet been systematically employed or centered in administrative studies. Consequently, noise persists as a largely "unseen error" in administrative decision-making, leaving promising reduction interventions unexplored. We use the disability-claims example to introduce noise as a field-wide problem, but our empirical investigation focuses on a single, well-defined predictive domain (public personnel selection) where the concept can be measured, and mitigation strategies tested.

[11]While classic literature in Public Administration (Frederickson, 1997; Musgrave, 1959) often aligns consistent treatment with notions of equity ('horizontal equity,' 'social equity'), we follow recent conceptualizations that define 'equality' as "treating like cases alike" and 'equity' as "treating dissimilar cases differently to achieve equal outcomes" (Raaphorst, 2021; Raaphorst et al., 2026).

Research Focus

This article introduces and explicates the concept of 'noise,' delineating its constituent components and establishing its relevance to PA. We aim to provide a systematic vocabulary and framework for addressing noise, filling a conceptual and analytical gap that currently exists in the field. Specifically, we demonstrate how noise, distinct from systematic bias, undermines administrative consistency and core public values: efficiency, equality, and legitimacy. This analysis underscores the imperative to broaden scholarly and practitioner attention to noise. Our work therefore proceeds in two parts. First, conceptually: How can the 'noise' framework be adapted for PA to analyze unwarranted decisional variability, and what are the implications for foundational public values? Second, we empirically investigate a promising noise-reduction avenue: the integration of artificial intelligence (AI) into administrative decision-making.

Distinguishing deterministic algorithms from AI applications is essential. *Algorithms* are well-defined computational procedures that produce identical outputs from specific inputs (e.g., Cormen et al., 2022, p. 4; Knuth, 1977, p. 63). In contrast, *AI applications* are complex systems that leverage techniques like machine learning and natural language processing to emulate cognitive functions and perform tasks traditionally requiring human intelligence (Glikson and Woolley, 2020, p. 631). While deterministic algorithms are, by definition, noise-free (identical Y for identical X), the consistency of AI applications varies. Nevertheless, many such systems, including machine learning scoring models as the AI tool we examine, offer significant noise-reduction potential, exhibiting substantially less variability than human decision-makers.

Specifically, we investigate two human-AI collaboration modes for mitigating noise in public personnel selection: (1) 'ensembling,' which involves the statistical aggregation of independent human and AI assessments into composite scores, and (2) behavioral 'specialization,' wherein humans render the final judgment after considering AI-provided advice. This yields evidence-based insights into leveraging AI for enhanced public sector consistency.

We focus on personnel selection as it exemplifies a consequential, uncertain, and noise-susceptible judgment domain. Selecting the right people is one of the most influential managerial decisions, affecting organizational performance (Luan et al., 2019) and, in the public sector, organizational legitimacy through mechanisms such as representative bureaucracy (Jakobsen et al., 2023). Despite the prevalence of formal evaluation procedures, the predictive validity of common selection methods remains limited (Sackett et al., 2022), and practitioners continue to rely heavily on subjective judgment (Highhouse, 2008). As Highhouse and Brooks (2023) argue, it is precisely the inconsistency in how different evaluators interpret and weight identical candidate evidence that proves particularly pernicious for judgment quality. This combination of formal structure and residual subjectivity makes personnel selection an appropriate setting for investigating noise: formal enough that consistency is expected, yet subjective enough that noise can flourish. Kahneman et al. (2021, p. 301) reach the same conclusion, identifying selection as a "perfect test case" because of its extensive research base, "offering lessons that can be extrapolated to many judgments involving a choice among several options." Moreover, personnel selection is increasingly a domain of AI deployment in public organizations, making it a timely context for evaluating human-AI collaboration modes. Despite this, PA research has devoted limited attention to personnel selection (Jakobsen et al., 2023), reinforcing the need for empirical

investigation. The unwarranted variability introduced by noise in this domain causes suboptimal hiring outcomes and creates unequal candidate opportunities, directly implicating administrative efficiency and equality.

Drawing on data from prior multi-study experiments (Keppeler et al., 2025), we utilize parallel assessments of public-sector candidates from human recruiters and an AI application. We evaluate both collaboration modes' capacity to mitigate noise, defined as the variability in assessment ratings. For ensembling, we compare the variance of recruiters' initial ratings against composite scores aggregating human and AI ratings at various weighted contributions (up to 50% AI weight). For specialization, where recruiters revised initial assessments after AI exposure, we compare the variance of ratings with and without AI consultation. Our analysis demonstrates that both modes substantially reduce noise. Specifically, ensembling achieves noise reduction proportional to the AI's weighted contribution, while specialization—exposure to AI advice—significantly decreases judgment variance in the subsequent human decision. While both approaches are effective, ensembling yields a more pronounced reduction, contingent on the weight assigned to the AI component.

Contribution to Public Administration

This article advances PA scholarship by integrating insights from three distinct domains: theories on 'noise' and 'unwanted variability' in human judgment (Brighton and Gigerenzer, 2015; Kahneman et al., 2021; Sunstein, 2022), scholarship on the design and deployment of public sector AI (Engstrom et al., 2020; Keppeler, 2024; Keppeler et al., 2025; Margetts and Dorobantu, 2019), and foundational PA literature on public values (Bozeman, 2007; Frederickson, 1997).

First, we advocate for systematically engaging with noise—unwanted judgmental variability that fundamentally undermines efficiency and equality in administrative decision-making. This call connects with, and seeks to reinvigorate, one of PA's foundational pillars: the pursuit of consistent, predictable, and rule-bound administrative action. For example, the ideal of consistency was central to Weber (1978)'s emphasis on calculability within the rational-legal model. Likewise, Wilson (1887) and White (1926) sought to establish principles for a more orderly and impartial public service. Noise directly erodes this bedrock ideal of consistent governance. While early reforms pursued consistency primarily through hierarchical controls, structures, and standardized procedures, the noise framework offers a contemporary analytical lens, particularly as AI presents complementary tools for achieving consistency. Relatedly, the prioritization of noise mitigation becomes particularly acute as PA confronts increasing administrative complexity, data-driven governance, and heightened demands for consistent service delivery. If decisional accuracy is a concern, then the measurement and mitigation of noise is essential. We aim to catalyze research identifying, measuring, and reducing noise through targeted interventions.

Second, we contribute to the burgeoning scholarship on AI in PA by documenting AI's frequently overlooked potential benefit: noise mitigation. Our empirical examination of ensembling and specialization in public personnel selection offers specific, testable approaches for enhancing decisional consistency. As part of this empirical contribution, we also introduce a novel metric, the 'Specialization-Ensemble Weighting Point' (SEWP), which quantifies the noise-reduction

efficacy of AI-assisted specialization as an equivalent percentage of AI weighting in an ensembling structure. This provides a standardized benchmark for comparing these distinct human-AI collaborative strategies.

Third, leveraging behavioral PA's experimental methodology (Bertelli and Riccucci, 2022; Grimmelikhuijsen et al., 2016; Moseley and Thomann, 2021), we demonstrate how re-analyzing existing experimental data enables direct noise assessment and intervention evaluation. This provides a practical blueprint for PA researchers to investigate noise and mitigation strategies using available datasets, thereby contributing to more reliable and equitable public decision-making.

7.2 The Concept of Noise

While 'noise' broadly refers to unwanted variability in the judgments of the same case (Brighton and Gigerenzer, 2015; Kahneman et al., 2021), its conceptual jurisdiction comprises distinct subcomponents. In line with Kahneman et al. (2021), noise can be decomposed into so-called 'level noise' and 'pattern noise'—with the latter including both 'stable pattern noise' and 'occasion noise.' *Level noise* refers to the variability in the average level of judgments by different decision-makers. For example, in administrative settings, some employees may be systematically harsher, while others are more lenient, when evaluating similar cases. In contrast, *pattern noise* represents the variability in decision-makers' responses to particular cases. Specifically, *stable pattern noise* refers to patterned differences between decision-makers regarding the influence of case characteristics and merits. For example, in personnel selection, one recruiter may prioritize a candidate's formal education, while another weighs professional experience more heavily, leading to divergent evaluations of the same applicant. Both level noise and stable pattern noise thus concern between-rater variability in judgments. In contrast, *occasion noise* is a component of pattern noise that relates to within-rater inconsistency; this occurs when a decision-maker evaluates functionally equivalent cases differently due to transient factors (such as fatigue or mood at the moment of decision-making).

Within this framework, administrative "error" in decision-making—outcomes that contradict established policies, rules, and public values—is an additive function of systematic bias and noise (i.e., Total Judgment Error = Bias + Noise). While *bias* refers to systematic differences in the average treatment of citizens and cases (e.g., discrimination based on race or gender), *noise* is the variability in decisions for the same (or functionally identical) citizens and cases, manifested via three pathways: between-rater *level noise*, between-rater *stable pattern noise*, and within-rater *occasion noise*. Also, noise is inherently context-dependent; much like systematic bias, its prevalence is expected to be more pronounced in domains and tasks characterized by high workloads, systemic uncertainty, and significant discretionary leeway.

The Problem of Noise

The fallibility and inherent variability of human judgment have long been acknowledged, from seminal work on 'bounded rationality' (Simon, 1947) to foundational studies demonstrating the superiority of statistical over clinical prediction in numerous domains (Meehl, 1954), and broader contributions from psychology and behavioral economics (Brunswik, 1955; Tversky and Kahneman, 1974; see also Hammond, 1996). It is well established that different individuals, or even the same individual at different times, may reach divergent conclusions when faced with ostensibly similar decision tasks. Indeed, such noise has been documented across numerous domains, including judicial sentencing (Chen and Loecher, 2025; Clancy et al., 1981; Danziger et al., 2011; Eren and Mocan, 2018; Kleinberg et al., 2018), forensic science (Ulery et al., 2012), university admissions (Simonsohn, 2007), researcher choices in research design and interpretation (Huntington-Klein et al., 2025; Menkveld et al., 2024), medical diagnostics (Beam et al., 1996), and asylum decisions (Chen and Loecher, 2025; Heyes and Saberian, 2019).

However, while such inconsistency is broadly acknowledged, the systematic study of noise has lacked the conceptual presence and focused research attention afforded its more conspicuous counterpart, bias. As Kahneman et al. (2021, pp. 5–6) emphasize: "*To understand error in judgment, we must understand both bias and noise [. . .] But in public conversations about human error and in organizations all over the world, noise is rarely recognized. Bias is the star of the show. Noise is a bit player, usually offstage.*" This prevalent tendency to overemphasize systematic bias while underappreciating the often substantial contribution of random variance to overall error constitutes what Brighton and Gigerenzer (2015) term the 'bias bias.'

PA scholarship has commendably invested substantial effort to identify and understand systemic biases, discrimination, disparities, and inequities within administrative decision-making and service delivery. This includes foundational work on 'street-level bureaucracy' (Lipsky, 1980) and social equity (Frederickson, 1997), 'administrative burdens' as sources of inequality (Herd and Moynihan, 2018), racial equity (Gooden, 2014), and a wealth of empirical studies documenting these challenges (Bell and Jilke, 2024; Guul, 2023; Guul et al., 2021; Hansen, 2024; Jilke and Tummers, 2018; Olsen et al., 2022; Pedersen and Nielsen, 2020; Pedersen et al., 2018; Pfaff et al., 2021; Riccucci and Van Ryzin, 2017)

Noise and Bias

Noise represents a fundamental source of error in human judgment. Far from necessarily being a lesser problem than systematic bias, noise is an independent and additive component of total judgmental error (Brighton and Gigerenzer, 2015; Kahneman et al., 2021). Consequently, its impact on the quality and fairness of decisions can be as damaging as—and in some contexts, more significant than—that of systematic bias.

The extent to which bias or noise constitutes the more critical driver of error in any specific administrative context depends on their respective magnitudes; a system may be predominantly noisy, predominantly biased, or suffer significantly from both. However, noise is almost always present and often large (Kahneman et al., 2021), yet it frequently goes unaddressed. Similarly, Brighton and Gigerenzer (2015) explain how the variance component (noise) often dominates total

judgmental error, particularly under conditions of high uncertainty where simple models that minimize variance often outperform more complex, but higher-variance, approaches. Importantly, for an individual citizen, a decision distorted by random variability can be as detrimental and unjust as one skewed by a systematic error.

Unlike systematic bias, which denotes systematic errors pointing in a predictable (albeit erroneous) direction, noise is the random, unpredictable dispersion of judgments around an average (which itself may also be biased). Thus, systematic bias and noise are conceptually distinct and orthogonal constructs; the presence or magnitude of one does not inherently predict the presence or magnitude of the other (Brighton and Gigerenzer, 2015; Kahneman et al., 2021). A system can be highly biased yet exhibit minimal noise (all judgments are consistently erroneous in the same manner), or it can be, on average, unbiased but suffer from extreme noise (judgments are widely scattered yet may still center on the correct value).

Figure 18 illustrates this distinction using the classic bullseye analogy, where the target's center represents the correct or accurate decision, and the dots depict individual judgments.

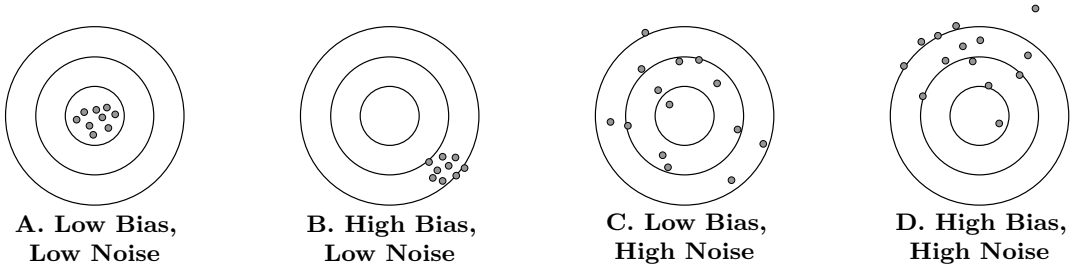


Figure 18: Bias and Noise

Notes: Illustration inspired by Kahneman et al. (2016). Bias is the average error, visualized as the offset of the judgment clusters' mean from the center. Noise is the variability within these judgments, visualized as their dispersion. The figure thus demonstrates the *orthogonality* of these two error sources: the location of the clusters' mean (bias) and the magnitude of their dispersion (noise) can vary independently.

As a concrete example, consider a panel of recruiters evaluating candidate applications for a public service role during an initial screening phase. Systematic bias might manifest if individual recruiters on the panel consistently tend to undervalue candidates from non-traditional educational backgrounds or overvalue a specific type of past experience detailed in their submitted materials. This would lead to systematically lower (or higher) scores for certain candidate profiles, irrespective of their true potential (as illustrated in Figures 18B or 18D). Noise, however, would be evident if individual recruiters on the panel assign disparate job fit evaluations to the identical candidate application, or if a given recruiter assesses objectively similar applications markedly differently across separate batches of applicant reviews. This latter manifestation of noise could arise from factors like the assessment of the immediately preceding application, the recruiter's current workload, or even their personal mood (see Danziger et al., 2011, for a judicial analogy). This judgmental inconsistency reflects the scatter observed in the high-noise targets (Figures 18C and 18D). In the hiring context, noise implies that a candidate's likelihood of being invited for an interview is unduly influenced by random, irrelevant factors, such as which specific recruiter reviews their application or even when that same recruiter reviews it. For further elaboration of the concepts illustrated by Figure 18, see Appendix E.1, Section I.

The Challenge to Public Values

For the individual citizen, the impact of noise in administrative decision-making extends beyond statistical nuances. As established in the preceding section, noise is a significant error source. Even in the absence of systematic bias, substantial noise *alone* can render decisions unreliable and unjust, thereby directly undermining core public values foundational to the legitimacy of government and public administration (Bozeman, 2007; Denhardt and Denhardt, 2015; Beck Jørgensen, 2007; Beck Jørgensen and Bozeman, 2007). When administrative decisions are undermined by noise—creating a 'lottery of outcomes' (Kahneman et al., 2021)—public values such as efficiency and equality are challenged.

Beyond detracting from consistency and predictability values, noise can directly compromise administrative *efficiency*. Inconsistent decision-making breeds inefficiency, for instance, by necessitating additional resources for reviews, appeals, and corrections of erroneous judgments (Kahneman et al., 2021). Furthermore, errors attributable to noise can lead to increased citizen inquiries, consuming administrative time and effort. The unpredictability inherent in noisy systems can also hinder optimal resource allocation and organizational learning, as the lack of consistent feedback loops, obscured by unwanted variability, impedes the identification of the true efficacy of policies and procedures.

The challenge that noise poses to fair and equal administrative decision-making is particularly acute. *Equality*, in this context, demands that like cases should be treated alike (Raaphorst, 2021; Raaphorst et al., 2026), a principle realized through the consistent application of rules (Frederickson, 1997). By definition, noise strongly contravenes this tenet by introducing arbitrary variability in outcomes for individuals who are, by all relevant criteria, similarly situated. This commitment to equal treatment is a foundational value in public administration, central to the pursuit of administrative justice and social equity (Frederickson, 1997).

For the individual citizen, receiving an unfairly harsh (or lenient) decision compared to similar others due to random noise is as detrimental as receiving one due to systematic bias (Sunstein, 2022), irrespective of whether system-wide averages appear acceptable. This arbitrary inconsistency from the perspective of the citizen signifies that access to public services, the severity of sanctions, or the allocation of benefits may depend on the 'luck of the draw' (Kahneman et al., 2021).

Furthermore, such random differentiation can erode both perceived and actual administrative justice and impartiality, as outcomes appear capricious rather than aligned with established policy rules and procedures. The presence of significant noise, therefore, inherently diminishes legitimacy, regardless of whether citizens perceive it. For further discussion of the legitimacy concerns, see Appendix E.1, Section II.

Addressing judgmental noise, therefore, is not merely a technical exercise for improving decisional accuracy but is a democratic imperative. It is integral to the principle of legality and the public value of equality—standards that citizens rightfully expect their administrative agencies to embody and defend (Bozeman, 2007; Wright, 2015).

7.3 Towards More Consistent Decision-Making

Fortunately, noise is not an immutable characteristic of human judgment and decision-making; it can be measured and managed. A potential initial step is a 'noise audit'—a systematic process designed to quantify the variability in judgments rendered by multiple professionals evaluating the same set of cases independently (Kahneman et al., 2016; Kahneman et al., 2021). The utility and feasibility of conducting such noise audits are increasingly recognized, even in complex domains such as the labeling of benchmarks (Kejriwal et al., 2024). Such audits can render the invisible problem of noise visible, compelling organizations to acknowledge its existence and impact.

Moreover, beyond diagnosis, a suite of 'decision hygiene' strategies can be implemented to mitigate noise. These strategies draw from an extensive tradition of research aimed at improving judgment by reducing the influence of irrelevant factors and structuring decision processes (Arkes, 1991; Larrick, 2004; Milkman et al., 2009). Key approaches include: (a) establishing clear rules and standards—recognizing the often superior performance of even simple actuarial (i.e., algorithmic/rule-based) methods over unaided human intuition (Dawes et al., 1989; Grove et al., 2000), partly because such rules can significantly reduce error variance (Brighton and Gigerenzer, 2015); (b) decomposing judgments into smaller, independently assessed components; (c) utilizing structured decision-making protocols; (d) employing relative judgments rather than absolute scales where appropriate; and (e) training decision-makers to enhance their awareness of sources of variability.

Aligning with Simon (1947)'s foundational work on administrative behavior, these strategies emphasize a broader point regarding the nature of noise in administrative decision-making: organizational structures, procedures, and environmental conditions directly structure its prevalence and scale. Relatedly, it is critical to note that focusing on noise does not suggest a lack of existing organizational structures or interventions that help to minimize decision inconsistency. Nor does it imply that public professionals are not oriented toward consistent decision-making. Rather, much like the established scholarship on systematic bias and discrimination, the 'noise agenda' calls for renewed and focused attention to the concept itself, its consequences, and its potential mitigation. Existing bureaucratic structures may already suppress noise (Bovens and Zouridis, 2002; Lipsky, 1980); our claim is not that safeguards are absent but that they are unevenly applied and rarely measured, leaving residual variability undiagnosed.

One of the most potent strategies for noise reduction is the aggregation of multiple independent judgments. This principle, closely related to the 'wisdom of crowds' (Surowiecki, 2004), posits that averaging several independent assessments tends to cancel out random individual errors, thereby yielding a more accurate and reliable collective judgment. The benefits of such aggregation are particularly pronounced when the individual predictors offer diverse, uncorrelated insights (Page, 2007). For instance, in personnel selection, having several recruiters independently evaluate each candidate and subsequently averaging their scores can significantly reduce the noise inherent in any single recruiter's assessment.

However, despite its theoretical appeal, this approach presents considerable practical challenges. First, it is resource-intensive—demanding the time, effort, and associated costs of multiple

decision-makers—and ensuring genuine independence among decision-makers can be difficult (Sunstein, 2005)[12]. Second, particularly in personnel selection, decision-makers often resist analytical methods, failing to appreciate the inherently probabilistic and error-prone nature of the task (Highhouse, 2008). Collectively, these limitations create a compelling case for exploring complementary strategies, such as the integration of AI.

We note that the problem of noise, as outlined in the preceding sections, is a fundamental issue inherent in human judgment, *independent* of any particular technological solution like AI. Indeed, the conceptualization of noise and many suggested strategies for its reduction predate the advent of AI. While this article investigates AI's potential for noise reduction, the imperative to address unwanted judgmental variability exists regardless of one's stance on AI.

Algorithms and AI

Technological advancements offer a potential remedy to the inherent challenges of human judgmental variability. Kahneman et al. (2016) identify the replacement of human judgment with algorithms—which use case-specific data to generate predictions or decisions—as "*the most radical solution to the noise problem*," citing their inherent advantage of being noise-free. Because a formula consistently yields the same output for any given input, even simple and imperfect algorithms can achieve superior consistency compared to human professionals (Kahneman et al., 2021). The impetus to explore such algorithmic solutions is strengthened by findings revealing the stubborn persistence of human judgmental noise, even in highly controlled settings. For example, significant noise has been documented among human experts tasked with meticulously labeling data (Kejriwal et al., 2024). Such evidence of noise in human-generated outputs, even when precision is paramount, underscores the challenge of achieving consistent human judgment and thus enhances the appeal of inherently consistent algorithms.

However, replacing human judgment with deterministic algorithms, while effective for noise reduction, may be operationally or politically untenable in certain public sector contexts (Kahneman et al., 2016; Sunstein, 2022). While a similar strategy of full replacement by AI systems—full AI 'automation'—may also prove infeasible or undesirable in certain public sector domains, human-AI collaborative frameworks (Hillebrand et al., 2025; Raisch and Krakowski, 2021) present a more pragmatic and promising alternative.

This potential is increasingly evident as the proliferation of AI propels a transition toward an algorithmic public sector (Bullock et al., 2020; Keppeler et al., 2025; Sidhu et al., 2024; Sousa et al., 2019), fundamentally reshaping policy implementation, administrative decision-making, and public service delivery (Bullock, 2019; Dunleavy and Margetts, 2023; Young et al., 2019). This transformation signifies a growing integration of AI into routine public agency operations, prompting calls to "rethink government" with AI (Margetts and Dorobantu, 2019; see also Coglianese and Lehr, 2017; Engstrom et al., 2020). Numerous opportunities are emerging, ranging from optimizing resource allocation and automating or augmenting routine tasks to identifying patterns and needs that human administrators might overlook. Such applications offer

[12]For example, assessors may be implicitly influenced by the opinions of others (even if expressed non-verbally), susceptible to prior information cascades, or swayed by conformity pressures within a group, thereby correlating their errors and diminishing the noise-reducing potential of aggregation.

potential benefits, including enhanced efficiency and fairness through data-driven optimization and standardized procedures (Busch and Henriksen, 2018; Chatterjee et al., 2022; Jarrahi, 2018; Van Noordt and Misuraca, 2022).

However, deploying AI in government is by no means without significant challenges and considerations. Scholars highlight attendant technical, ethical, and democratic risks, rightfully cautioning that the integration of AI in the public sector can erode public agency legitimacy (Busuioc, 2021; Grimmelikhuijsen and Meijer, 2022; Kuziemski and Misuraca, 2020; Schiff et al., 2022; Wirtz et al., 2020). For further review of prominent and legitimate concerns, see Appendix E.1, Section III.

Thus, achieving a more complete perspective on the benefits and drawbacks of AI in the public sector requires dedicated PA research into AI's implications for noise in administrative decision-making. While scholarship has yet to pinpoint the exact decision types where AI tools serve as particularly effective means of mitigation, consequential judgment domains characterized by high uncertainty and broad discretion (such as hiring) exemplify noise-susceptible settings appropriate for empirical testing. Importantly, AI applications can vary considerably in their output consistency. While generative AI tools (e.g., Large Language Models) may exhibit substantive *occasion noise*, other AI tools (like machine learning scoring models such as the one that we examine in the present study) often share the consistency advantage of deterministic algorithms[13].

Human-AI Collaboration

Full automation, wherein AI systems independently render and implement decisions without human intervention, is potentially efficient for highly structured, low-risk tasks. Importantly, however, its deployment in complex administrative decisions affecting citizens also elicits significant concerns regarding discretion, fairness, and due process—challenges that reflect the feasibility issues broadly identified by Kahneman et al. (2016) and Kahneman et al. (2021). Consequently, and often more appropriately for many public administration contexts, AI is typically employed through human-AI collaboration modes (Raisch and Krakowski, 2021; Hillebrand et al., 2025) with the aim of synergistically leveraging the distinct strengths of both human judgment (e.g., contextual understanding, ethical reasoning, adaptability) and AI capabilities (e.g., processing vast datasets, consistency, recognizing complex patterns). Indeed, recent meta-analytic evidence underscores the utility of such human-AI combinations over human-only systems (Vaccaro et al., 2024).

Broadly, these modes of human-AI collaboration encompass 'human-in-the-loop' models, where humans actively interact with and guide (or train) the AI during the decision process (Slade et al., 2024), and 'gatekeeping' models, where humans oversee, approve, or exercise veto power over AI-generated recommendations (Rahwan, 2018). These arrangements are

[13]Traditional machine learning models, such as those used for scoring or matching, typically produce identical outputs from identical inputs once trained, exhibiting very low or zero *occasion noise*—much like the deterministic algorithms. Generative AI models, by contrast, can produce variable responses to identical prompts depending on parameter settings (e.g., temperature). The AI tool examined in this study belongs to the former category: it is a machine learning scoring model, and in the present study it produced a consistent person-job fit score for each candidate.

situated within a broader spectrum of human-AI collaboration, representing varying levels of AI authority; however, they are uniformly designed to *augment* rather than merely *automate* human tasks (Agrawal et al., 2018; Akata et al., 2020; Hillebrand et al., 2025; Raisch and Krakowski, 2021). Our research, however, focuses on two distinct human-AI collaborative modes—ensembling and specialization—that are particularly pertinent to noise reduction in administrative decision-making.

Ensembling and Specialization. As previously discussed, the inherent consistency of well-designed algorithms offers a compelling approach for mitigating noise (Sunstein, 2022). Because a deterministic algorithm produces identical output from identical input, it eliminates any contribution to 'occasion noise.' It is important to note, however, that not all AI applications share this absolute consistency. As elaborated later, generative AI models can—much like humans—produce variable responses to identical prompts. Overall, however, AI applications—and machine learning models in particular—are likely to exhibit substantially less variability than human judges, and therefore hold significant potential for noise reduction. This article concentrates on two collaborative frameworks where this potential is particularly salient:

Ensembling: This mode directly applies the principle of aggregating independent judgments—with 'ensembling' broadly defined as the "aggregation of predictions" from multiple sources concerning the same issue (Choudhary et al., 2025)—but with a crucial enhancement: one of the "judges" is an AI. This approach differs from other scholarly distinctions of AI in decision-making (e.g., full automation versus 'human-in-the-loop' or 'gate-keeping' augmentation), as the AI here functions more as an equivalent decision agent alongside the human, with both addressing the identical prediction problem in its entirety (Keppeler et al., 2025; Puranam, 2021). In the context of public personnel selection, an ensembling strategy could involve a human recruiter providing an assessment of a candidate's job fit, while an AI—trained on relevant data, such as curricula vitae, assessment results, and historical performance metrics—independently generates its own assessment. These two (or more) assessments are then mechanically aggregated, through simple averaging or a weighted formula, to produce a final composite score (Silver, 2012; Tetlock and Gardner, 2015)—a method proven effective for combining judgments (Soll and Larrick, 2009). The underlying rationale is that the AI's more consistent judgment (Kahneman et al., 2021)—though potentially susceptible to systematic bias (Sunstein, 2022)—when combined with the human's (potentially noisy yet context-aware) judgment, will yield a final decision that is more reliable and less prone to the errors of either party acting in isolation. The mechanical aggregation ensures that the AI's consistent signal directly contributes to reducing the overall noise in final decisions.

Specialization: This mode, aligning more closely with traditional augmentation perspectives, involves AI and humans performing distinct, specialized subtasks, wherein the AI primarily functions as an advice-giver. The AI does not directly contribute to a mechanically aggregated score but rather serves as a decision support tool for the human decision-maker (Bonaccio and Dalal, 2006). In the context of public personnel selection, for instance, an AI might analyze application materials and provide the human recruiter with a predicted job fit score, highlight key qualifications, or flag potential areas of concern. The human recruiter then considers this AI-generated advice alongside their own assessment before rendering the final hiring decision.

Critically, the practical efficacy of specialization may necessitate that the human agent is explicitly incentivized, mandated, or prompted to consider the AI's advice.

The potential for noise reduction through specialization is less direct than that offered by ensembling; its efficacy hinges fundamentally on the human recruiter's responsiveness to, and appropriate assimilation of, the AI's advice. Recent research highlights that human-AI feedback loops can alter human judgments (Glickman and Sharot, 2025); however, individuals may exhibit 'selective adherence' to AI advice, particularly when it corroborates their initial assessments (Alon-Barkat and Busuioc, 2023; Selten et al., 2023). If the AI furnishes a consistent, high-quality signal, and the human decision-maker utilizes this signal to anchor their judgments or to critically evaluate their initial impressions, this can prompt them to apply criteria more consistently across candidates and over time. This, in turn, can reduce their individual 'occasion noise' and potentially foster closer alignment with a shared standard, thereby mitigating both 'occasion noise' and 'system noise.' Conversely, if recruiters disregard the AI, systematically discount its advice (Burton et al., 2020; Dietvorst et al., 2015), or over-rely on it without critical scrutiny—tendencies that encompass both over- and under-reliance (Vaccaro et al., 2024) as well as 'algorithm appreciation' in certain contexts (Logg et al., 2019)—the noise-reduction benefits may not materialize.

Expected Noise Reduction. We anticipate that both ensembling and specialization—as two modes of human-AI collaboration—can contribute to reducing noise, including within the context of public personnel selection. In both modes, the AI's contribution can serve as an anchor, dampening the random fluctuations inherent in human assessments. However, ensembling, owing to its mechanical aggregation of a noise-free AI judgment with human judgment, may yield a more substantial reduction in overall judgment variability compared to the specialization mode, wherein the human recruiter retains full discretion.

Personnel selection is a predictive judgment: recruiters estimate an applicant's future person–job fit, a quantity that is, in principle, verifiable against a criterion such as subsequent performance. This classification matters because the noise-reduction tools we deploy are precisely those the noise literature reserves for predictive tasks. Mechanical aggregation (ensembling) and the disciplined use of a low-variance external signal to anchor human assessment (specialization) presuppose that judgments track an underlying true value; their justification is strongest where a criterion exists and weakest for purely evaluative judgments, i.e. matters of value or contestable preference, where convergence on a single score may be neither feasible nor normatively desirable.

Diagnostic judgments, which likewise admit a verifiable answer, pattern with predictive ones in this respect. Predictive (and diagnostic) judgments dominate consequential administrative tasks such as eligibility determination, risk scoring, and selection, which is why decision-hygiene strategies and algorithmic anchoring travel well across these domains. Recognizing this scope condition sharpens rather than narrows the contribution: it specifies where human–AI collaboration should reduce noise—criterion-bearing predictive tasks—and flags where its appropriateness is an open normative question. We therefore treat personnel selection as a paradigmatic predictive judgment (for a comparable approach, see Highhouse and Brooks, 2023) and the predictive/evaluative distinction as the boundary condition governing our mitigation strategies.

Nevertheless, specialization, where the AI functions as an advisor, may represent a more common, practically feasible, and acceptable approach in many public sector contexts, aligning with the established augmentation paradigm (Daugherty and Wilson, 2018; Kuziemski and Misuraca, 2020; Raisch and Krakowski, 2021). For further discussion of these implementation challenges, see Appendix E.1, Section IV.

Drawing from the conceptualization of noise in public sector decision-making, and to empirically investigate AI's potential for noise reduction in public personnel selection, this study examines how human-AI collaboration may affect noise in the assessment of job candidates. Specifically, our hypothesis is as follows:

H1: The *specialization* mode is associated with lower noise (i.e., reduced variance in job fit ratings for similar candidates) relative to human-only decision-making.

Moreover, we examine the noise reduction efficacy of the *ensembling mode* as an analytical benchmark. As the mechanical aggregation of a high-consistency AI rating with noisy human ratings will axiomatically reduce the variance of the pooled ratings by construction, we refrain from stating a formal hypothesis on the ensembling-noise relationship. However, our ensembling focus and findings provide a quantification of the mode's noise reduction potential and serve as a mechanical reference point for evaluating the behavioral specialization result.

Furthermore, we directly examine the relative noise-reduction efficacy of specialization and ensembling. While ensembling is theoretically expected to yield a greater effect due to its mechanical nature, we refrain from positing a formal hypothesis directly comparing the two modes. This is because the efficacy of ensembling is not a fixed value but is instead entirely contingent upon the specific weighting assigned to the AI's input; a 50% AI weight, for instance, will necessarily be more effective than a 10% weight. Moreover, no existing theory provides a basis for predicting the specific weighting at which ensembling would produce a noise reduction equivalent to that achieved through specialization. Consequently, rather than formulating a direct comparative hypothesis, our analytical approach involves exploring the effect of ensembling across a continuous spectrum of AI weights (from >0% to 50%). This methodology allows for a more nuanced comparison and enables us to empirically identify and discuss the ensembling weight that corresponds to the noise reduction achieved through specialization.

7.4 Data and Design

This study uses data from two prior experiments that investigated the deployment and disclosure of AI in public personnel selection (Keppeler, 2024; Keppeler et al., 2025), hereafter designated Original Study 1 and Original Study 2. The following subsections outline the core components of the experimental designs and the data pertinent to the present analysis. Comprehensive details of the original studies are provided in the primary publications (Keppeler, 2024; Keppeler et al., 2025). The data and replication code for the current study are publicly available on the Harvard Dataverse.

Data Source: Prior Experimental Studies

Original Study 1 (OS1) employed a proprietary, commercially established AI tool for direct sourcing (talentwunder.com) in a live hiring initiative for three hard-to-fill positions within a German public organization. We recognize that the tool's proprietary and commercial nature raises "black box" questions about architectural transparency and version stability. However, its use in a live hiring initiative provides high ecological validity, reflecting the types of external solutions currently procured by public agencies. Moreover, consistent with our analytical focus, our findings rely on the AI's inherent consistency rather than its specific internal logic. Insofar as algorithmic systems typically exhibit significantly lower noise than human raters, the fundamental principles of noise reduction identified here should be theoretically stable across diverse platforms.

The AI tool identified 2,000 potentially suitable candidates by sourcing profiles from Xing, a German professional networking platform analogous to LinkedIn. For each candidate, the platform's machine learning model generated a person-job fit score (on a 0-100% scale) based on a textual analysis of the job descriptions and the candidate profiles. Specifically, this machine learning AI model—distinct from generative AI systems—produced a single, consistent score for each candidate's profile (i.e., the AI's scores signify zero within-candidate variance). This property is fundamental to the inferences of our analytical results. While the proprietary nature of the tool limits full transparency into its model architecture and training data, our empirical findings refer to consistent AI output for identical candidate data.

Original Study 2 (OS2) was a survey experiment involving a sample of 538 German managers with hiring responsibilities, who were recruited via public sector job advertisements between January and September 2023. The study was administered online; participants completed the tasks asynchronously and without monitoring or time constraints to ensure a naturalistic decision environment. The procedure unfolded in several stages. First, participants reviewed a real job advertisement from OS1 (Finance Application Specialist). Second, they evaluated the person-job fit (0-100%) of four candidate profiles selected from the OS1 pool, which were systematically varied by gender and ethnicity. To prevent consistency artifacts, participants were not exposed to the same or highly similar cases repeatedly. Third, following these initial ratings, participants were randomly assigned to receive identical advisory job fit scores for each candidate. The critical manipulation was the framing of these scores—the actual AI ratings from OS1—as originating from either "an AI" or "a human peer" (a manager from a recruitment department). Finally, participants could revise their initial assessments considering this advice.

Current Study: Data and Operationalization

As illustrated in Table 6, the present analysis operationalizes our hypothesis and the ensembling benchmark by integrating data from both OS1 and OS2.

Decision Mode	**Description / Operationalization**	**Data Source(s)**
Human-Only Baseline	The variance in participants' initial, unaided person-job fit ratings, representing the level of judgmental inconsistency (noise)	Initial human ratings (OS2)

Decision Mode	Description / Operationalization	Data Source(s)
Specialization	The effect of AI advice on judgment, measured by the change in variance from initial to revised (post-advice) ratings	Initial & Revised human ratings (OS2)
Ensembling	A weighted statistical average of a participant's initial rating and the AI's advisory score[1]	Initial human ratings (OS2) & AI ratings (OS1)

Table 6: Operationalization of Decision-Making Modes

Notes: The baseline represents independent human judgment. Specialization represents a sequential revision process where the human retains final judgment, while ensembling represents a mechanical integration of independent judgments. [1] The ensembling variable was calculated as: $Ensemble\ Score = ((1\text{-}w) \cdot HumanRating) + (w \cdot AIRating)$, where w represents the weight assigned to the AI rating.

Human-Only Baseline: Our baseline measure of human decision-making is defined as the participants' initial person-job fit ratings for the four candidate profiles, representing judgments made prior to receiving any advice.

Specialization Mode: To test the *specialization* framework, we compare the variance across participants' initial ratings (within each profile) with the variance in their revised (post-advice) ratings. This direct comparison allows us to assess how AI-generated advice, when used as a decision support tool, influences the consistency of human judgment.

Ensembling Mode: The *ensembling* framework is operationalized by computationally constructing a series of composite scores. Each score represents a weighted average of a participant's initial human rating and the corresponding AI rating. To model a spectrum of integration strategies, the weight assigned to the AI's rating was systematically varied from >0% to 50%. This simulation enables a direct comparison between the variance of the initial human-only ratings and the variance of the resulting ensembled scores at each weight increment, thereby quantifying the noise-reduction efficacy of ensembling across a continuum of AI influence.

Our primary analyses are restricted to the OS2 subsample of participants who were randomly assigned to the AI-advice condition. This approach enables a direct and robust comparison between the specialization and ensembling modes, as both are anchored to the same initial human judgment from a single cohort. The full OS2 sample, which includes the human-peer advice condition, is reserved for robustness checks and auxiliary analyses to assess the effect of advice source framing on noise. Sample characteristics are detailed in Appendix E.2.

Analytical Strategies

Consistent with Kahneman et al. (2021)'s notion of system noise, we operationalize decision noise as the variance (standard deviation) of person-job fit ratings across respondents, computed within each of the four candidate profiles and then averaged. A lower average within-candidate standard deviation (σ) signifies reduced noise and thus greater judgmental consistency.

Our main analytical strategy involves comparing the variance across three distinct rating conditions: (1) *Human-Only Baseline*: The initial ratings provided by participants before exposure to any advice, (2) *Specialization Mode*: The revised ratings provided by participants

after exposure to AI-framed advice, and (3) *Ensembling Mode*: The composite scores derived from weighted averages of human and AI ratings.

This comparative framework allows us to quantify the noise reduction attributable to each mode of human-AI collaboration. To test *H1*, which posits that the *specialization* mode reduces noise, we compare the variance of the pre-advice ratings to that of the post-advice ratings. We employ two complementary statistical approaches to assess this change. First, we estimate 95% confidence intervals for the standard deviations of both rating sets using a clustered bootstrapping procedure (1,000 resamples). Clustering is performed at the respondent level to account for the non-independence of the multiple ratings provided by each participant. The 95% confidence intervals are derived from the 2.5th and 97.5th percentiles of the resulting bootstrap distribution. Second, to complement the bootstrap analysis, we conduct a formal statistical test for the equality of variances using Levene's test, which evaluates the null hypothesis (H_0) that the variances of the pre-advice and post-advice ratings are equal.

For the ensembling benchmark, we compare the variance of the initial human-only ratings against the variance of the simulated ensembled scores. Although our analysis models the effect across a continuous spectrum of AI weights (from >0% to 50%), for reporting clarity, we present the results at five key increments of AI contribution: 10%, 20%, 30%, 40%, and 50%. For each of these five comparisons against the baseline, we evaluate the change in variance using the same statistical procedures employed for the specialization analysis: (a) constructing 95% confidence intervals for the standard deviations via clustered bootstrapping and (b) conducting Levene's test for the equality of variances (H_0: equal variances).

Finally, to directly compare the noise-reduction efficacy of the two collaborative modes, we benchmark the performance of the specialization strategy against the ensembling framework. To formalize this comparison, we introduce the 'Specialization-Ensemble Weighting Point' (SEWP), a metric that quantifies the noise-reduction effect of specialization in terms of an equivalent AI weight within an ensembling framework.

Conceptually, the SEWP represents the percentage of AI contribution in a hypothetical ensembled score required to achieve the same level of noise reduction observed in the specialization mode. Generally, the equivalence point can be estimated and visualized by locating the post-advice variance on a continuum of potential ensembled variances. This estimation method is applicable in scenarios where the AI's output itself exhibits variability. Conversely, in cases where the AI's output variance approximates zero, the SEWP can be calculated directly as the proportional reduction in the standard deviation from the pre-advice (human-only) to the post-advice (specialization) ratings:

$$\text{SEWP} = 1 - \frac{\text{SD}_{\text{post}}}{\text{SD}_{\text{pre}}} \tag{22}$$

The full derivation of the SEWP is provided in Appendix E.3. A SEWP of 0 indicates no noise reduction from specialization, while a SEWP of 1 signifies complete judgmental convergence, equivalent to all decision-makers adopting a single, noise-free rating. This metric thus provides a direct and intuitive benchmark for comparing the two modes and allows us to report the effect of specialization using the common currency of an ensemble weight.

7.5 Results

This section evaluates the potential for ensembling and specialization modes of human-AI collaboration to reduce noise in public personnel selection. We begin with a descriptive overview of the findings, including a key preliminary analysis of behavioral responsiveness to AI advice. This is followed by the formal statistical test of our hypothesis and a comparative analysis of the two modes' effectiveness, concluding with robustness and auxiliary analyses.

Behavioral Responsiveness to AI Advice

Before testing our hypothesis, it is essential to establish a behavioral precondition for the specialization mode: that human decision-makers are responsive to AI advice. While ensembling is a mechanical strategy, the noise-reduction efficacy of specialization depends on participants actively incorporating the AI's input.

Our analysis reveals that 43.2% of participants adjusted their job fit ratings after receiving the AI's input. While the majority (56.8%) made no change—a finding consistent with similar 'assess-advise-reassess' experimental designs (e.g., Kennedy et al., 2022)—a substantial minority actively engaged with the advice. Critically, among those who did adjust their ratings, an overwhelming majority (96.2%) moved their assessments *closer* to the AI's score. This finding confirms that participants were broadly receptive to the AI's input, thereby validating a core assumption of the specialization framework. Within this group of "assimilators," adjustments were nearly balanced between those who decreased (52.4%) and increased (47.6%) their initial scores, with a mean absolute change of 13.7 points (SD = 9.07) on the 0–100 scale.

Descriptive Overview of Rating Distributions

Figure 19 presents the distributions of the key rating outcomes central to our analysis. The initial human-only ratings exhibit considerable dispersion across the 0-100 scale, a clear visualization of substantial baseline noise. In stark contrast, the distributions for both the post-advice ratings (specialization mode) and the AI-weighted composite scores (ensembling mode) are visibly more concentrated around their central tendency. This narrowing of the distributions provides strong preliminary evidence that both specialization and ensembling are associated with a marked reduction in judgmental variability, lending visual support to our hypothesis (and to the ensembling benchmark).

Analysis of Noise Reduction

To formally quantify the patterns observed in Figure 19, Table 7 details the descriptive statistics for each key rating condition. Figure 20 complements this table by visually plotting both the mean and the standard deviation for each condition, with the latter—our primary measure of noise—displayed alongside its corresponding bootstrapped 95% confidence interval.

Specialization Reduces Noise. As illustrated in Table 7 and Figure 20, our analysis of the

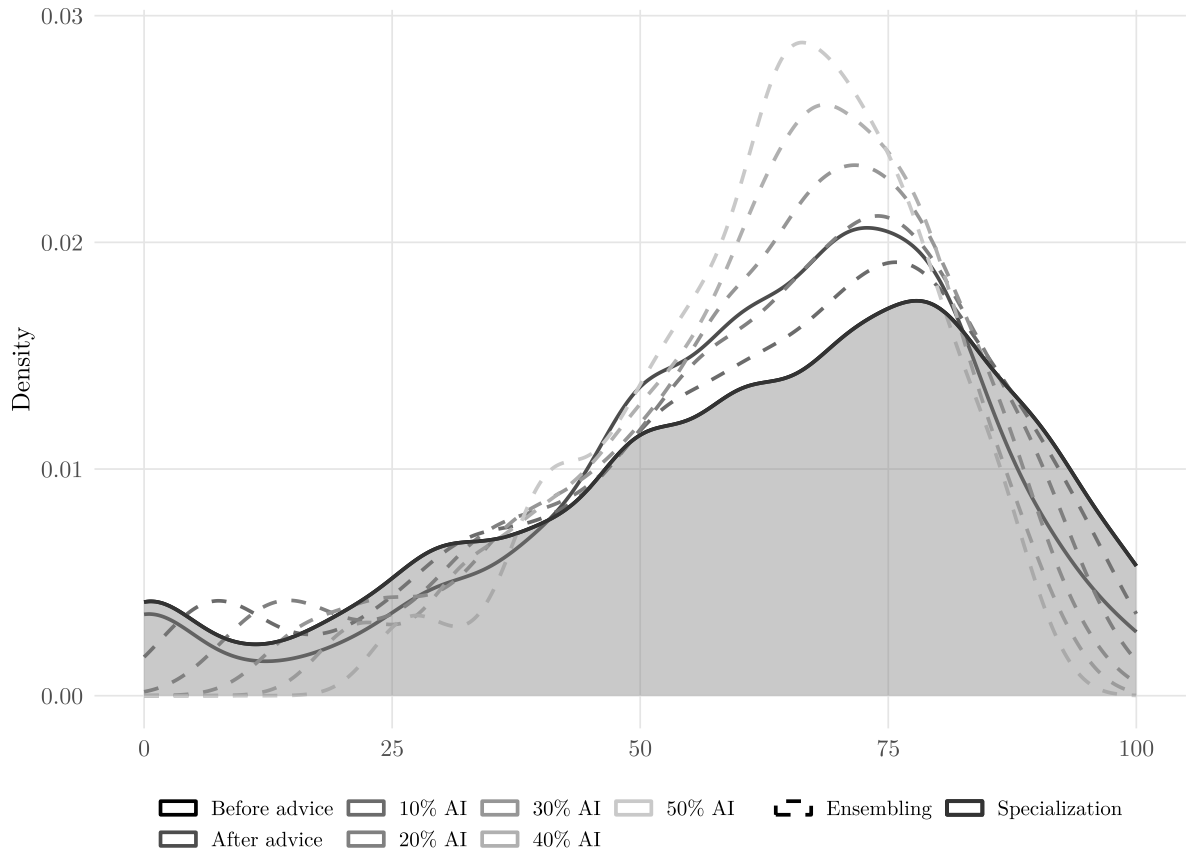


Figure 19: Plot of Outcome Distributions

Notes: Density plots illustrating the distributions of key rating outcomes. The shaded area represents the baseline human-only (pre-advice) ratings. Also shown are distributions for the post-advice ratings (specialization) and for ensembled scores at five levels of AI contribution (10%, 20%, 30%, 40%, and 50%). The x-axis indicates the job fit rating on a 0-100 scale.

	Mean (M)	SD (σ)	Variance (σ^2)	Min	Max	CI (lower)	CI (upper)
Human-Only Baseline (pre-advice)	61.09	25.41	645.5	0.00	100.00	59.56	62.62
Specialization Mode (post-advice)	60.93	22.06	486.7	0.00	100.00	59.59	62.28
Ensembling Mode							
10% AI	61.55	22.87	522.8	4.90	98.00	60.17	62.93
20% AI	62.02	20.32	413.1	9.80	96.00	60.78	63.25
30% AI	62.49	17.78	316.3	14.70	94.00	61.39	63.58
40% AI	62.95	15.24	232.4	19.60	92.00	61.98	63.93
50% AI	63.42	12.70	161.4	24.50	90.00	62.55	64.29

Table 7: Descriptive Statistics of Job Fit Ratings by Decision Mode

Notes: N = 1,084 (ratings) from 271 respondents. Shows the mean (M) and its 95% confidence interval (CI), the standard deviation (SD), variance, and range (min/max).

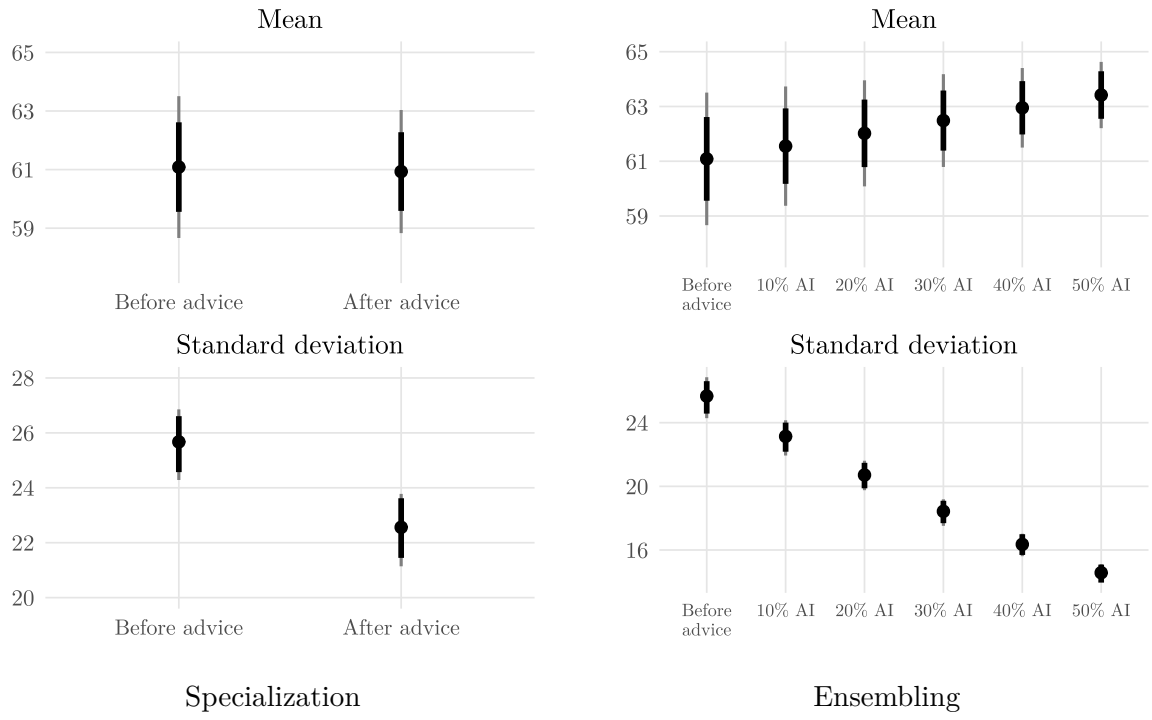


Figure 20: Mean Job Fit Ratings and Standard Deviations by Decision Mode

Notes: This figure illustrates the effect of the specialization and ensembling modes on the mean (top panels) and standard deviation (bottom panels) of job fit ratings. The left panel (Specialization) compares the pre-advice ratings with the post-advice ratings. The right panel (Ensembling) compares the pre-advice ratings with a series of ensembled scores, plotted as a function of the AI's weight in the composite score (in 10-percentage-point increments from 10% to 50%). Vertical lines represent two types of bootstrapped 95% confidence intervals: unclustered (thicker lines) and clustered at the respondent level (thinner lines). Clustering accounts for the non-independence of ratings provided by the same participant.

specialization mode reveals two key findings. First, exposure to AI advice did not systematically shift the average rating; the mean job fit score post-advice ($M = 60.93$) was not significantly different from the pre-advice average ($M = 61.09$; $p = .88$). Beyond this non-significant result, a formal equivalence test confirmed that the shift in ratings was statistically negligible, representing a trivial effect[14].

Second, we observe a substantial and statistically significant reduction in judgmental noise. The standard deviation of ratings fell from 25.41 (Variance = 645.5) in the human-only condition to 22.06 (Variance = 486.7) after participants received AI advice. This constitutes a 13.2% reduction in the standard deviation (a 3.3-point drop), indicating a marked increase in judgmental consistency. This reduction in variance is statistically significant, as confirmed by Levene's test ($p < .001$; see Appendix E.4).

Therefore, these results provide strong support for H1: the specialization mode significantly reduces noise without altering the average judgment of the raters.

Ensembling Mechanically Reduces Noise. Turning to the ensembling mode, we first address its effect on the mean rating before examining our primary outcome of interest, noise. Unlike in the specialization mode, the mean of the ensembled scores increases systematically with the AI's weighting (see Figure 20). This upward shift is an expected mathematical artifact, not a behavioral finding; it occurs because the proprietary AI's average rating (M = 65.75) was higher than the participants' initial average (M = 61.09). Consequently, incorporating the AI's higher ratings into a weighted average mechanically pulls the mean upward, a shift that becomes statistically significant when the AI's weight reaches 39.5% ($p < .05$)[15].

The central finding, however, pertains to the substantial reduction in noise. While variance reduction is inherent to ensembling, this reduction becomes statistically detectable (at $p < .05$) once the AI's weight in the composite score reaches 6.5%. The analysis reveals a dose-response relationship as expected: the noise-reduction effect becomes more pronounced as the AI's contribution increases. For instance, an ensembled score with a 25% AI weight reduces the initial standard deviation by 6.4 percentage points, while a 50% AI weight yields an 11.1-point drop. We can also show this statistically with Levene's tests, confirming the statistical significance of these reductions at each reported increment ($p < .01$ for 10% AI weight; $p < .001$ for 20% to 50% weights; see Appendix E.4).

Thus, the ensembling mode is a powerful and mechanically reliable method for mitigating noise in public personnel selection[16].

[14]We performed a Two One-Sided Tests (TOST) procedure (Lakens, 2017). To provide a conservative estimate that accounts for the nested nature of the data (1,084 ratings within 271 respondents), we conducted the test based on the number of independent participants ($N = 271$). Using an equivalence bound of Cohen's $d = \pm 0.15$, the observed difference was significantly within the equivalence bounds, $t(270) = 2.32, p = .011$.

[15]This aggregate effect on the mean was not uniform across the four candidate profiles. A profile-level analysis (see Appendix E.5, Figure A2) reveals that the mean ensembled score decreased for one candidate, increased for two, and showed no significant change for the fourth. This variance underscores the mechanical nature of the effect: the direction of the mean shift for any individual profile depends on whether the AI's rating for that specific profile was higher or lower than the human average.

[16]This mechanical shrinkage could in principle reflect that the AI's judgment was coarser than human judgment—collapsing distinctions between candidates—rather than more consistent. Empirically, however, the AI's scores for the four profiles (49, 55, 79, 80) yield a between-candidate standard deviation of 16.1, whereas the mean of respondents' within-rater between-candidate standard deviations is 4.4. While coarser judgment would compress rating differences between candidates, the AI tool distinguishes *more* across candidates than our human raters, not less.

Relative Efficacy of Specialization and Ensembling. Having established that both modes reduce noise, we now turn to their relative efficacy. To benchmark the behavioral effect of specialization against the mechanical effect of ensembling, we calculated the SEWP.

As visualized in Figure 21, our analysis yields a SEWP of 13.2% (95% CI: [10.3%, 16.4%])[17]. This result has a direct interpretation: the noise reduction achieved through the behavioral specialization mode is statistically equivalent to that of a mechanical ensembling strategy where the AI's rating is given a 13.2% weight. This finding provides a clear benchmark for comparison. It implies that an ensembling strategy with an AI weight greater than the upper bound of our confidence interval (16.4%) will likely reliably outperform specialization in reducing noise. Conversely, an AI weight below the lower bound (10.3%) would likely be less effective. More broadly, our results demonstrate that even a relatively modest AI weight (e.g., 15-20%) within a mechanical ensembling framework can match or exceed the noise reduction attained through a more hands-off, advice-giving approach.

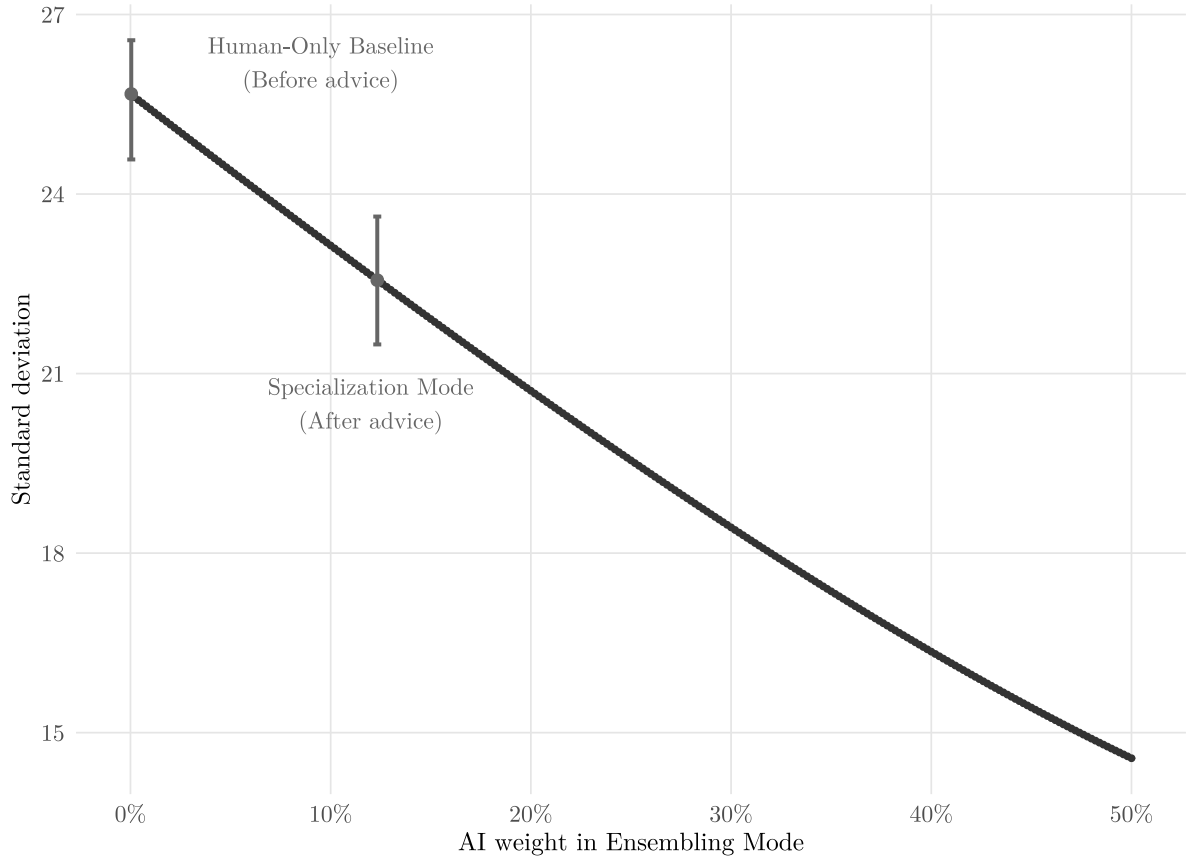


Figure 21: Comparing the Noise-Reduction Efficacy of Specialization and Ensembling

Notes: This figure illustrates the comparative noise-reduction efficacy of the ensembling and specialization modes. The y-axis represents decision noise, operationalized as the standard deviation of job fit ratings. The downward-trending line shows the standard deviation for the ensembling mode as a function of the AI's weight in the composite score (x-axis), ranging from 0% (the human-only baseline) to 50%. The separate point indicates the standard deviation achieved by the behavioral specialization mode. Vertical error bars on all points represent bootstrapped 95% confidence intervals. The figure provides a direct visualization of the Specialization-Ensemble Weighting Point (SEWP); projecting the specialization noise level onto the ensembling curve reveals the equivalent AI ensembling weight on the x-axis.

[17]We calculate the SEWP using Equation (22). We construct the 95% CI by substituting the 2.5th and 97.5th percentiles from a bootstrapped distribution of the standard deviation for the SD_{post} term in the same equation.

Beyond its analytical utility, the SEWP offers practical value for navigating AI adoption in the public sector. A specialization strategy might be more organizationally palatable because it aligns with the status quo and preserves human autonomy and discretion. In contrast, an ensembling approach that grants AI more direct decision-making authority may represent a more radical and less palatable shift, as human autonomy and discretion are key drivers of technology acceptance (Dietvorst et al., 2018; Ingrams et al., 2022). In these common scenarios, the SEWP serves as a critical evaluative tool. It (a) quantifies the impact of AI advice by providing a tangible measure of the noise reduction achieved; (b) aids in the justification of such systems by translating the behavioral effect into a clear, communicable metric; and (c) informs strategic choice by offering a data-driven benchmark to help managers decide whether the noise reduction from an advisory system is sufficient.

7.6 Robustness and Auxiliary Analyses

To validate the robustness of our findings and explore supplementary insights, we conduct a series of robustness and auxiliary analyses. First, we disaggregate the data to analyze the robustness of our findings by each of the four candidate profiles. Overall, as detailed in Appendix E.5, both the main noise-reduction effects and the SEWP appear relatively robust across profiles and are not driven by any single case.

Second, we assess the robustness of our findings using the full sample data. As our primary analyses are restricted to the AI-advice subsample of OS2, we replicate these analyses on the full OS2 sample, which also includes the human-peer advice condition. As detailed in Appendix E.6, these replications fully confirm our results. We find that the noise-reduction efficacy of specialization did not depend on the perceived source of the advice. Aligning with recent research indicating that citizens hold public organizations similarly responsible for discrimination in public services, whether it is algorithmic or human-driven (Alon-Barkat et al., 2025), we observe no significant difference in rating variance for participants who believed the advice came from "an AI" versus those who believed it came from "a human peer." Consequently, this comparable efficacy indicates that AI can serve as a viable—and likely more cost-effective—substitute for human peer advice when the primary objective is to enhance judgmental consistency.

7.7 Discussion and Conclusion

This article has pursued a dual objective: first, to introduce the critical but under-researched concept of 'noise' into the mainstream of PA scholarship, and second, to investigate the potential of human-AI collaboration to mitigate this source of decisional error. We established the theoretical framework by delineating the concept of noise—an "unseen error" in PA—and its relationship to systematic bias. By highlighting noise as a substantive, measurable threat to public values like efficiency and equality, this article calls for a renewed scholarly focus on achieving consistency in public organizations and administrative decision-making. While the concept of noise and its consequences are currently less recognized by scholars, decision-makers,

and the broader public than that of systematic bias, its detrimental impact on core public values warrants greater cognizance and urgent attention.

Building on this conceptual foundation, our study provided empirical proof-of-concept on how human-AI collaboration can secure more consistent administrative processes. Through a secondary analysis of experimental data in public personnel selection, we demonstrated how both ensembling (aggregation of human and AI ratings) and specialization (AI-as-advisor) strategies significantly reduce noise. A further contribution is the introduction of the 'Specialization-Ensemble Weighting Point' (SEWP), a novel metric that quantifies the efficacy of the specialization approach in the currency of an ensembling weight, thereby enabling a direct comparison between these distinct collaborative modes.

Importance for PA

Our conceptual arguments and empirical findings offer four contributions to the field of PA. First, this article issues a call for PA scholarship to systematically engage with the problem of noise. We emphasize that noise is not merely a theoretical construct but a tangible threat to constitutive public values. The continued neglect of noise means that efforts to improve administrative decision-making, which concentrate on bias, address only one dimension of total judgment error. Our work thus seeks to catalyze a research agenda within PA—one that moves beyond merely identifying noise to actively developing and rigorously evaluating reduction strategies. To reduce errors in administrative decision-making, noise must be measured and mitigated with the same rigor as systematic bias. Our findings provide empirical support for this agenda, demonstrating that AI can significantly enhance judgmental consistency (i.e., reduce noise) in public sector decision-making. This is a critical contribution, as more consistent decisions promote greater equality and, ultimately, justice in the public sector (Binns, 2022). Consistency, however, is necessary but not sufficient for fair decision-making: a process can be perfectly consistent and still be systematically wrong. Reducing noise is therefore a foundational *component* of judgmental quality, not a substitute for addressing bias.

Second, this research contributes to the burgeoning PA literature on AI by documenting a critical benefit: its capacity for noise mitigation. While much of the discourse has rightly focused on AI's inherent risks, such as algorithmic bias, its ability to enforce consistency and thereby reduce random error is a potential advantage. Our findings on ensembling and specialization provide concrete, testable models for how human-AI collaboration can enhance decisional reliability in high-stakes domains like public personnel selection. In doing so, we advance the conversation from generalities about AI's promise or peril to specific, evidence-based applications within public administrative practice.

An alternative approach for mitigating noise is the systematic assignment of multiple human assessors to a given decision task. Aligning with the intuition of 'the wisdom of crowds' (Surowiecki, 2004), aggregating independent, better-than-random human judgments reduces variance-based error. Indeed, across many personnel selection settings, such practice is already manifested in interview committees. However, this procedure is also cost-intensive, imposing strain on human resources and efficiency values. Therefore, human-AI collaboration modes

may constitute cost-effective venues for noise reduction: ensembling may essentially offer an extra (synthetic) decision-maker, while specialization may improve decision consistency, thus providing a cost-effective substitute for an additional human reviewer. Our finding that AI advice's noise reduction was equivalent to that of human-peer advice (Appendix E.6) underpins this perspective.

Third, this article offers a methodological contribution by providing a practical and resource-efficient template for investigating administrative noise through the repurposing of existing experimental data. Our approach—quantifying judgmental variance and using parallel AI assessments to examine noise-reduction strategies—enables scholars to study decisional consistency without undertaking costly new data collection. This replicable model for evaluating interventions like ensembling and specialization thus expands the scope for impactful research into improving the reliability of public decision-making.

Fourth, while our findings highlight AI's potential for noise reduction, they also raise the critical issue of whether this gain is achieved without introducing or shifting systematic bias. Our results indicate that specialization and ensembling may have distinct implications. In the *specialization* mode, the stability of the average rating post-advice indicates that the 'AI-as-advisor' introduced no new systematic bias at the aggregate level. For those concerned about flawed AI models, this positions specialization as a safer approach for reducing noise without risking the introduction of algorithmic bias. The *ensembling* mode, however, shifted the mean score by incorporating the AI's higher average rating, leaving it an open empirical question whether this shift moves judgments toward or away from greater accuracy. This ambiguity speaks to a central tension in the AI debate: proponents might believe that the AI's assessment could be more accurate and correct for human biases, while critics could contend the opposite.

A pressing concern—echoing the concept of the bias-variance trade-off in machine learning—is that reduction in noise may (re)introduce bias, and vice versa. Some fixes for noise, like rigid guidelines or AI-usage, may entrench bias, while some fixes for bias may birth new sources of noise. The bias-variance trade-off, however, is not a theorem. As supported by our specialization results, noise may be mitigated without introducing bias. Accordingly, an intervention may simultaneously mitigate both bias and noise (e.g., Axt and Lai, 2019). Important questions thus arise: What characterizes the contexts where a bias-variance trade-off arises? Which interventions can effectively mitigate both bias and noise?

Ultimately, this underscores the imperative to evaluate administrative systems on both noise and bias simultaneously. By design, our work isolates and focuses on noise, thereby complementing the rich PA literature on bias. A key takeaway is that neither component of error should be ignored; fair and accurate administrative decision-making requires a holistic approach that concurrently measures and manages both unwanted variability (noise) and systematic deviation (bias).

Methodological Limitations

While this study offers novel insights, some methodological limitations warrant caution and provide clear directions for future research. As detailed in Appendix E.7, our findings are best

understood as proof-of-concept and should be tempered by important considerations regarding (a) the artificial nature of our survey-based research design, (b) the potential for broader external generalizability (e.g., to decision domains other than personnel selection), (c) the need for comprehensive evaluation that integrates measures of noise, bias, and overall error, and (d) the properties and quality of the specific AI tool (e.g., the proprietary tool's "black box" nature and potential for version updates may influence the observed effects and limit exact cross-platform replication).

Perspectives for Future Research

This study opens several promising avenues for future inquiry within PA. First, scholars should quantify the magnitude and identify the primary drivers of noise across diverse public sector domains, from regulatory enforcement and benefit allocation to social service delivery. Key questions include: What is the typical magnitude of noise in these areas, and what are its primary drivers? Based on such mapping, future research is warranted to develop a more granular theory of noise in PA, encompassing also the specific types of administrative decisions where noise prevalence and scale are most acute, to identify the most appropriate mitigation strategies for each. Similarly, while our findings document the noise reduction efficacy of human-AI collaboration, future research should seek to decompose the mechanisms, isolating reductions in 'level noise' from 'stable pattern noise' and 'occasion noise.'

Second, future research should develop and test a wider array of noise-reduction interventions. Beyond further investigation into 'decision hygiene' techniques, this includes refining human-AI collaboration modes. Promising areas include exploring dynamic weighting in ensembling, assessing the impact of AI explainability on advice-taking in specialization, and designing more sophisticated 'human-in-the-loop' systems (Slade et al., 2024). A critical line of inquiry for ensembling is to model how its efficacy changes when the AI component itself is biased (cf. Sunstein, 2022). More generally, research should also seek to pinpoint the underlying mechanisms of noise reduction interventions. For example, does mitigation occur through more deliberate processing or through more implicit responses to social conformity pressures?

Third, our findings and SEWP metric require replication and validation—in alignment with the call for increased replication efforts within PA (Pedersen and Stritch, 2018; Walker et al., 2019). Replicating this study in different contexts is key for identifying the scope conditions of our results and confirming the validity and reliability of our inferences.

Fourth, the shared correlates between noise and bias remain a critical frontier. Research should investigate whether interventions targeting one inadvertently affect the other and explore strategies that can mitigate both simultaneously. Understanding the psychological mechanisms of human-AI interaction, including how long-term feedback loops alter human judgment (Glickman and Sharot, 2025), is fundamental to this effort and to designing effective and ethically sound collaborative systems. Relatedly, future research should map the public's 'noise tolerance' and compare these attitudes to perceptions of systematic bias in administrative decision-making.

Fifth, the durability of noise-reduction effects in specialization modes warrants investigation. While our study captured immediate adjustments, longitudinal research is needed to determine

whether repeated exposure to AI advice leads to sustained improvements in human judgment or if the effects are merely transient.

Finally, PA scholarship must continue to scrutinize the normative and ethical dimensions of using AI for noise reduction. This requires weighing the benefits of enhanced consistency against critical issues of accountability and transparency (Fine Licht and Fine Licht, 2020; Grimmelikhuijsen, 2023), due process, and the potential for unintended consequences (Sunstein, 2022), ensuring that these systems ultimately serve public values and human flourishing (Floridi et al., 2018).

Implications for Practice

This research offers key implications for public managers and policymakers. The first imperative is to recognize and audit for noise. Organizations should challenge the default assumption of judgmental consistency, particularly in decision-making contexts characterized by high uncertainty, subjective evaluation, or ambiguous criteria—conditions common from social service delivery to regulatory enforcement. As Kahneman et al. (2021) advocate, making this invisible problem visible is the necessary first step toward corrective action. Overlooking noise risks misattributing errors to systematic bias when random inconsistency may be the primary driver of poor outcomes (Brighton and Gigerenzer, 2015).

Second, our findings demonstrate that public managers can leverage AI as a powerful tool for enhancing consistency. Like deterministic algorithms, AI can be a valuable collaboration partner in creating more consistent administrative processes, but adoption must be cautious, evidence-based, and context-sensitive. We recommend that managers consider pilots to test different human-AI collaboration modes within their specific operational environments.

The choice between collaborative strategies like ensembling and specialization requires navigating distinct trade-offs. Ensembling offers potent, reliable noise reduction, but its utility is greatly contingent on the quality of the underlying AI model. Specialization, conversely, may be more organizationally and politically palatable, as it preserves more human decision autonomy and discretion (Dietvorst et al., 2018; Ingrams et al., 2022). Crucially, this strategic choice must extend beyond efficacy to encompass practical feasibility, organizational acceptance, and the human element. Practitioners must evaluate how different AI features might influence acceptance among both staff and the public. A key consideration is the human response to aggregation; managers must anticipate how professionals' behavior might change when they know their judgment is being combined with an AI's.

Furthermore, successful implementation hinges on establishing clear operational and training protocols. Managers must define how final scores will be applied, for example, whether an ensembled score automatically triggers an action (e.g., selecting interview candidates) or serves as a recommendation to a human who retains final authority, ideally with a requirement to document any deviation. Training protocols must foster not only the ability but also the motivation for professionals to interact effectively with AI outputs.

Ultimately, a key objective of public administration is to provide services that are legitimate, legal, effective, and equitable. Enhancing judgmental consistency—that is, reducing noise—is

paramount to this mission, as it is a direct means of promoting fairness and equality. By bringing the problem of noise into the PA discourse and demonstrating the potential of human-AI collaboration to address it, this article aims to spur both scholarly inquiry and practical reforms, thereby advancing the efficiency, equality, and legitimacy of public service.







Chapter 8

Longer Laws, Better Service? Legislative Complexity and Bureaucratic Quality

Co-authored with Jonas Krogh Madsen

Abstract

Most advanced democracies face accelerating growth in rules and regulation, but there is little empirical evidence on how this growth affects the bureaucracy. We ask: Do longer and more complex laws undermine or improve implementation? Prevailing theories often assume that legislative complexity burdens bureaucrats and reduces service quality. We revisit this assumption by linking paragraph-level changes in Danish social policy laws to administrative errors, measured through reviewed citizen complaints. Using over 112,000 complaints across 95 legal provisions from 2013–2024, we use within-paragraph variation in complexity, length, and cross-references. A typical increase in paragraph length is associated with an error rate about 0.5 percentage points lower. A typical increase in cross-referencing is associated with an error rate about 0.8 percentage points lower. The total number of complaints shows no comparable association with these legal changes. Our study highlights a caveat for simplification policy agendas: while complexity may appear inefficient, it can enhance rule clarity and promote more consistent service delivery.

8.1 Introduction

Does more complex legislation lead to worse implementation of public policies? Across many developed countries, governments are trying to simplify regulation. Using phrases such as better regulation, cutting red-tape, bureaucratic simplification, policies that are meant to stop the growth of legislation are popular across party lines and across different institutional regimes. As laws and regulations across advanced democracies have undeniably become increasingly extensive in terms of length, amendments, and regulatory embeddedness (Katz et al., 2020; Jakobsen and Mortensen, 2015; Hinterleitner et al., 2024), the predominant answer from scholars is yes: policy growth brings rising legal complexity that challenges bureaucratic capacity and undermines implementation quality (Adam, 2025; Young et al., 2024). For instance, Fernández-i-Marín et al. (2024), Gratton et al. (2021), and Giommoni et al. (2026) all argue that legal complexity complicates case management and leaves the bureaucracy overburdened. From this perspective, the layering of amendments to existing regulation, a dynamic described as "rules breed rules" (Katz et al., 2020; Jakobsen and Mortensen, 2015; Hinterleitner et al., 2024), both wastes effort and harms implementation.

Yet there is surprisingly little direct evidence that more elaborate legal provisions produce worse administrative decisions. Existing studies generally measure complexity or errors at the level of whole laws and at the level of whole countries or organizations. They can therefore not easily connect a change in legal rules to the quality of decisions made under that rule.

While the clustering of rules and rising legislative complexity might complicate implementation, rules are typically drafted with a purpose, e.g. to adapt to changing circumstances, to correct recurring mistakes, or to close gaps identified in practice. For instance, Ash et al. (2024) show that the size of a U.S. city's legal code scales with the size of the jurisdiction. In this sense, the accumulation of rules can be read as a way to reduce opportunities for divergent readings of the law, thereby lowering the risk of administrative errors, as Huber and Shipan (2002) argue. While most studies argue that ambiguity is bad, complexity could thus also entail specificity through which longer or denser legislation may actually reduce ambiguity, streamline interpretation, and support more consistent frontline implementation (Osnabrügge and Vannoni, 2025). Rather than viewing complexity as dysfunction, we argue that what is normally regarded as complexity can function as legislative specificity, deliberately crafted to guide bureaucrats and reduce erroneous decisions. On this account, added clauses, conditions, and cross-references serve to sharpen vague concepts, specify procedures and exceptions, and close loopholes that previously led to confusion and inconsistent interpretation. Complexity here is a means of clarification.

At the heart of this puzzle are competing assumptions about bureaucratic discretion. Discretion can be seen in two ways. On the one hand, it lets caseworkers adapt rules to complex individual circumstances and exercise professional judgment. On the other hand, it opens room for inconsistent interpretations, unequal treatment, and legal errors. Whether legislative complexity improves or impairs implementation depends on how it interacts with discretion. More complex laws can overload bureaucratic discretion by raising cognitive demands and make it harder to see what is relevant in a specific case, but they can also structure discretion by

specifying conditions and procedures more precisely, thereby narrowing the space for error-prone judgment.

We therefore frame our contribution in terms of two competing hypotheses: a "complexity-as-red tape" hypothesis, which expects increases in legal length, syntactic density, and interdependence to make it harder for bureaucrats to apply the law correctly and thereby produce more administrative errors; and a "complexity-as-clarification" hypothesis, which expects the opposite, as more detailed and interconnected laws should reduce the likelihood of error by constraining discretion and reducing interpretive ambiguity. Both mechanisms are theoretically plausible, but existing evidence is rarely able to test them directly at the point where these mechanisms should be found, namely between concrete legal provisions and the legality of the decisions made by bureaucrats in the frontline.

One very important reason is that administrative errors are hard to observe. This is not straightforward as laws tend to apply across a wide array of cases, and implementation rarely leaves a clear audit trail. Most administrative decisions, both at higher and lower levels of bureaucracy, are not reviewed, making it difficult to assess when and why errors occur. Moreover, legal texts often co-determine outcomes, requiring interpretation across multiple rules or provisions. As a result, much of the existing work relies on e.g. audits or aggregate measures of improper payments that indicate whether benefits have been awarded incorrectly at the program or agency level (Greer and Bullock, 2018; Ban and Hill, 2025; Young et al., 2024). Others do not measure bureaucratic errors at all, but simply treat the growth and density of law as a symptom of poor lawmaking and infer harm from outcomes further downstream (Gratton et al., 2021). These data, included in prominent studies of complexity, errors, and administrative performance in social programs, provide important information about overall error rates, but are generally not linked directly to specific legal provisions. Consequently, they offer little leverage on the causal mechanisms invoked by theories of legal complexity: whether changes in wording, length, or legal structure meaningfully alter patterns of bureaucratic error.

To overcome these empirical challenges, we turn to a unique setting, Danish municipal service delivery in social policy. In Denmark, municipalities implement centrally determined social policies, and their decisions are subject to appeal by citizens. When a citizen files a complaint and the municipality upholds its decision, the case can be escalated to the National Social Appeals Board. There, a panel assesses whether the original decision was valid or erroneous, effectively providing a quasi-audit of administrative quality. Crucially, what makes this setting unique is that these complaints are logged with a reference to the exact paragraph (or a small cluster of paragraphs) of the law in question, allowing us to observe not just whether an error occurred, but which legal provision governed the decision. As the underlying substance of the provision does not change, we can look at within-paragraph variation in legal complexity over time.

This structure allows us to empirically link specific statutory paragraphs directly to service quality outcomes across time. We are able to test whether complexity, measured through legislative length, syntactic depth, and textual features such as delegation clauses or cross-referencing, is associated with higher or lower rates of administrative error. We exploit rich variation in the legal provisions covering complaints across Denmark's 98 municipalities and

over 90 paragraphs of three major social laws, each governing services such as sickness benefits, rehabilitation, and support for vulnerable groups. We thus measure the correlation between the laws and errors at a very fine-grained within-paragraph level. By linking each citizen appeal to the exact statutory paragraph it contests and to a third-party reversal decision, we observe administrative error at the level of the individual legal provision.

Contrary to widespread assumptions, we find that complexity, measured by statutory length and cross-referencing, is associated with fewer administrative errors. An average increase in paragraph length is associated with a 0.5pp (corresponding to 1.23%) lower rate of administrative errors. An average increase in the use of cross-references is associated with a 0.7pp lower error rate, or about 1.8% of the baseline error rate.

Our findings thereby challenge the prevailing view that complexity uniformly degrades administrative performance. Instead, they suggest that longer and more elaborate laws may reduce discretionary ambiguity and clarify implementation, thus improving consistency. This resonates with findings by Osnabrügge and Vannoni (2025), who show that vague legal language is associated with weaker compliance across EU member states, and Ash et al. (2025), who find that more detailed legislation is associated with greater legal clarity and, through that, with higher economic growth.

8.2 Theory

Modern democracies experience continual policy growth as governments – by and large – produce new laws and regulations far more frequently than they remove outdated ones. Over time, this expansion results in an increasingly voluminous legal landscape both in terms of length and embeddedness (Hurka et al., 2022; Katz et al., 2020). Much scholarship interprets this development as a challenge for bureaucratic capacity as well as policy implementation (Fernández-i-Marín et al., 2024; Gratton et al., 2021): legal complexity exacerbates interpretive demands and raises the likelihood that both higher-level bureaucrats and, even more so, frontline bureaucrats at the street level will misapply or misunderstand the legal basis of their decisions (Young et al., 2024; Fernández-i-Marín et al., 2024). Yet, complexity may also structure discretion and support more consistent administrative practice. In this section, we outline these competing perspectives and develop expectations for how legislative complexity should shape bureaucratic error.

Policy growth, legal complexity, and administrative burden Several accounts explain why legal codes and policy portfolios keep growing. Governments respond to citizens' demand for action across an ever-wider range of problems (Adam et al., 2019); organised interests with institutional access shape what gets legislated (Ehrlich, 2011); and politicians' incompetence can also fuel legal growth (Gratton et al., 2021). Where these accounts are tested against one another, organised interests emerge as the principal driver and electoral competition looks largely irrelevant (Fernández-i-Marín et al., 2025). Because it is easier to add new rules than to repeal what already exists, policy accumulates (Hinterleitner et al., 2024; Limberg et al., 2021), at times into entangled, self-reinforcing "kludges" (Kawai et al., 2018). On a more functional reading, however, the same growth can be seen more as a structural response to an increasingly complex

society (Ash et al., 2024) and an expanding welfare and tax state (Katz et al., 2020).

A broad literature links the accumulation of laws to rising legal complexity. As governments rarely repeal insufficient rules, but rather add new policies that refer to existing legislation, legislative frameworks become increasingly layered, fragmented, and cross-referenced (Senninger, 2023; Adam et al., 2019), thereby producing denser and internally more interdependent bodies of law, characterized both by increased length and by more intricate formulations, detailed conditions, and procedural thresholds (Hinterleitner et al., 2024).

From this perspective, complexity is seen to impair the interpretive and operational capacity of bureaucracies and thereby challenge effective implementation. The core mechanism linking this complexity to administrative errors runs through the red tape it places on bureaucratic decision-making (Bozeman, 1993). As statutes grow, bureaucrats must process more information, integrate more conditions, and cross-check a larger set of provisions to ensure the legal basis of each decision. As a consequence, even small amendments to existing laws can require them to re-learn how different parts of the law interact, raising the cognitive demands of each decision. This increased cognitive load heightens the probability of overlooking relevant exceptions, misinterpreting thresholds, or applying the wrong provisions, even though substantive rules are clear. Moreover, these risks are only enhanced whenever bureaucrats face time pressures, incomplete information, or insufficient legal guidance. Administrative errors, then, become a product of the mismatch between demands imposed by complex legislation and the resources available to the bureaucrat. This not only matters for e.g. implementation willingness, burnout, and work motivation (Steijn and Voet, 2019; Tummers and Bekkers, 2014) but also invites cognitive and behavioural coping strategies such as simplification, avoidance, or reliance on heuristics (Davis and Pandey, 2024; Fuenzalida et al., 2026) that ultimately concern bureaucratic discretion. As a very core feature of bureaucratic decision-making, it allows bureaucrats to tailor decisions to complex circumstances, but when legal texts are dense, fragmented, or ambiguous, the very possibility to balance decisions with professional concerns also becomes a source of interpretive drift. In short, ambiguity within or across legal provisions expands the range of plausible readings, making complexity a source of error, not necessarily because bureaucrats are disobedient to rules, but because of the increasing likelihood that appellate authorities will simply deem interpretations incorrect.

Taken together, the "complexity-as-red tape" perspective links complexity to errors by asserting that complexity increases cognitive and procedural demands on bureaucrats, paves the way for unintended discretionary interpretations, and thereby increases the likelihood of administrative errors.

Legal specificity as clarification Yet, rather than viewing complexity solely as a cue for poor legislation (Gratton et al., 2021), we echo recent advances (see e.g. Senninger, 2023) that raise important methodological concerns with the approach that focuses too narrowly on e.g. readability as a proxy for complexity. Readability and complexity may not necessarily capture the same underlying construct. Laws could for instance be easy to read yet vague, under-specified, or conceptually thin, while long provisions may reflect deliberate attempts to tighten definitions, clarify exceptions, or codify procedures. Reforms building on this approach risk

undermining the ways in which detailed legal specification might provide more precise guidance for bureaucratic decision-making. Solving complexity with readable and simpler provisions may indeed reduce bureaucrats' cognitive load in the short term but increase discretionary ambiguity and inconsistency in the long term. As Katz and Bommarito (2014) caution, the aim should be to make law as simple as possible but not simpler, since under-specified rules that fail to reach their intended ends carry social costs of their own. In this perspective, legal complexity should not be seen as an impediment to implementation but as a potential instrument for improving rule clarity. Longer or more detailed laws may include specific conditions, definitions, and procedural steps that reduce ambiguity, standardize decision-making, and thereby ensure more equity in bureaucratic decision-making across bureaucrats (Huber and Shipan, 2002). This argument resonates with an older theoretical tradition that treats law as a tool for constraining arbitrary power and enabling consistent rule application (Fuller, 1969). Rather than assuming that e.g. length equals complexity, this approach distinguishes between complexity that is chaotic and complexity that is structured. Legal texts that elaborate rules in detail may appear complex by word count or syntactic depth, but they may in fact be deliberately designed to reduce interpretive variation among bureaucrats both at the government level and at street level. Underpinning this line of argument is also a different view of discretion as primarily problematic whenever rules are vague. Discretion is error-prone, not when rules are over-specified, but when they do not properly specify what counts as relevant information, how criteria should be interpreted, or how exceptions should be handled. For bureaucrats implementing policies in the frontline, it is a shortage of specification that makes them fill gaps with e.g. local norms or informal heuristics, which results in inconsistent decisions across contexts. Structured legal complexity thereby constrains this very discretionary space and produces more even and legally correct application of rules. Recent empirical advances recognize this distinction. For instance, Osnabrügge and Vannoni (2025) find that EU legislation with vague formulations and, to some extent, syntactic complexity is associated with lower compliance across member states, suggesting that underspecified law undermines correct implementation. Similarly, Vannoni et al. (2021) show that with increasing bureaucratic independence, legislators become more prone to respond by drafting more detailed statutes, which shows how complexity is used deliberately to reduce bureaucratic drift (further examples include Senninger (2023) and Kosti (2024)).

This "complexity-as-clarification" perspective thus argues that amendments that add detail, clarify definitions, or embed rules in existing legislation scaffold discretion by outlining extensive decision procedures and narrow the room for interpretations that differ from those appellate authorities would deem correct. The absence of specificity, not the presence of complexity, is what compromises implementation.

Competing perspectives and theoretical expectations These contrasting perspectives generate fundamentally different expectations about the relationship between legal complexity and administrative errors. The "complexity-as-red tape" view holds that longer and more interdependent statutes overwhelm administrative capacity, impose cognitive strain on bureaucrats, expand unintended discretion both within and across units, and increase the likelihood that frontline interpretations diverge from those applied by oversight bodies. The opposing

"complexity-as-clarification" view instead sees additional statutory complexity as a means to narrow this very space for ambiguity and discipline discretion to provide clearer guidance, reduce interpretive variance across bureaucrats, and push decision-making towards what is deemed correct by oversight bodies.

These juxtaposed perspectives therefore raise several empirical questions for future research: When does legal detail increase interpretive clarity rather than confusion? How do bureaucrats experience and navigate long or dense laws in everyday decision-making? And under what conditions does structured complexity support more consistent service delivery? Answering these questions in detail requires moving beyond surface metrics of complexity and attending to how legal texts structure discretion and action on the ground.

It is precisely along these lines that our study contributes. Focusing on a unique dataset from Danish social policy, we are able to link paragraph-level changes in legal texts, measured by, for example, sentence length, LIX, and cross-references, directly to paragraph-level variation in administrative errors identified by third-party national oversight authorities. In the next section, we describe how uniquely structured citizen complaint data allow us to trace changes in law to administrative outcomes at the level of individual legal paragraphs.

To examine this question, we turn to a case where the effects of legal complexity on implementation quality can be directly assessed: Danish social policy.

8.3 Empirical context and data

We study the relationship between changes in legislative complexity and the quality of public service delivery in Denmark between 2013 and 2024. Denmark is consistently ranked as having one of the highest levels of bureaucratic capacity across various measures (Grøn and Salomonsen, 2020). Thus, theoretically, discretion should be more beneficial here than in states with lower capacity because strong professional norms and credible control and review mechanisms are already in place, making discretion less risky (Fukuyama, 2013).

We focus on the implementation of Danish social policy, because this setting provides us with variation in the length and structure of specific legal provisions and allows us to plausibly connect this variation to observable administrative outcomes. Specifically, we examine three major legal frameworks: the Social Assistance Act (Lov om aktiv socialpolitik), the Social Services Act (Lov om social service), and the Sickness Benefits Act (Lov om sygedagpenge). These laws govern key areas of Danish welfare administration, including means-tested income support, assistance for individuals with physical or mental impairments, and temporary income replacement for the sick. Together, these programs account for a substantial share of municipal budgets and represent a core domain of frontline welfare provision.

The laws change often. For research design purposes, however, the legal architecture allows us to measure the same provision over time. When legislators introduce major substantive changes—such as new entitlements or eligibility criteria—these are typically written in the law as new paragraphs, whereas minor or procedural amendments are often incorporated into existing provisions. This modular legal structure means that we can get fine-grained tracking of legal

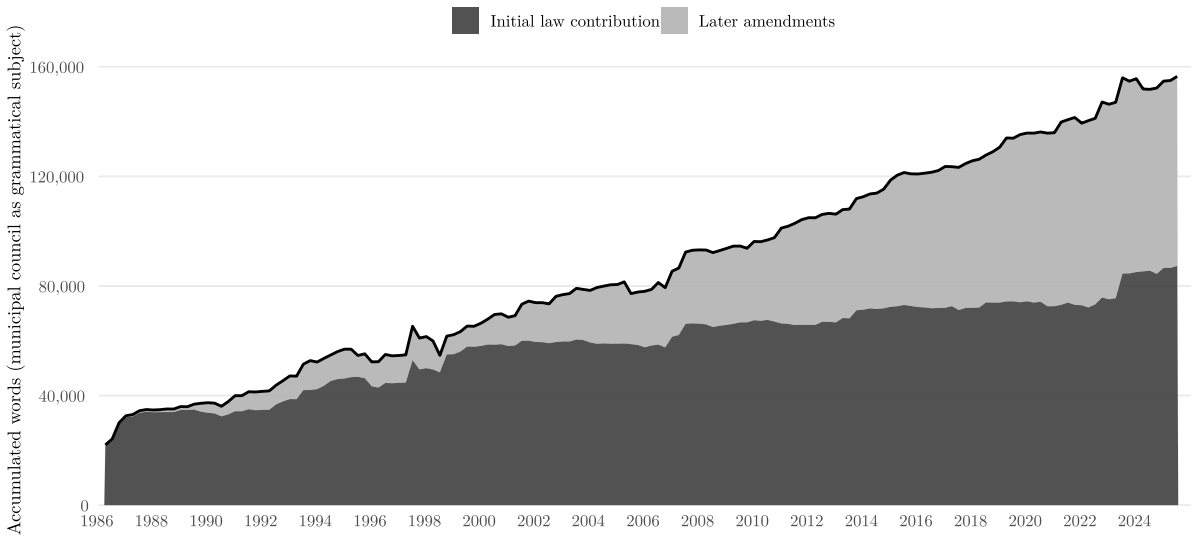


Figure 22: Accumulation of laws regulating Danish municipalities

Note: Accumulated statutory provisions that assign obligations or permissions to Danish municipalities, divided into initial law text and text added through later amendments. Source: retsinformation.dk.

changes at the paragraph level over time. While all three laws are enacted at the national level, they are implemented by caseworkers in the 98 municipalities, creating variation in how legal rules are interpreted and applied across local governments.

The body of law governing Danish municipalities has grown markedly over recent decades. However, much of this growth is not because many new laws have been written, but rather because many laws have been amended and made longer. Legislators often revise existing statutes rather than only passing new ones, changing provisions already in place. Figure 22 shows the resulting accumulation, and it is this ongoing revision of standing provisions that generates the within-paragraph variation our design exploits.

Administrative errors are generally difficult to observe in data. Existing studies have mostly relied on program audits, such as in Young et al. (2024). While audits provide valuable information about error rates, they are also infrequent, measured at the organizational level, and not tied to specific legal provisions. We study a setting in which service quality is continually audited: citizen complaints. Danish citizens can appeal municipal decisions to a national oversight body, the National Social Appeals Board (Ankestyrelsen), which determines whether the legal basis for the original decision of the municipality was correctly interpreted and applied. These decisions are then categorized according to whether the municipal judgment is upheld, reversed, or remanded. We therefore treat this as a way to measure administrative error, since if the Board overturns or remands a decision, it reflects a procedural or substantive flaw in local implementation.

Each appeal is linked to specific paragraphs of the law on which the original decision was based, allowing us to systematically connect legal complexity to administrative outcomes. Because these appeals are resolved under national standards, they also enable meaningful comparison across municipalities. In sum, this setting provides us with a panel structure containing granular legal and administrative data on 95 paragraphs across 46 quarters over 11 years, from 2013 to

2024. With this, we can assess how legal variation at the paragraph level—measured through changes in complexity—relates to observable patterns of service delivery and administrative error.

As some paragraphs are lumped together into other case topics, we exclude appeals that are mapped to case topics spanning more than 15 paragraphs. Most case topics, however, cover only one or two paragraphs. The list of included case topics with the corresponding paragraph(s) is in subsection F.1. Furthermore, not all paragraphs are relevant throughout the entire duration of the data set, since some of them are removed or added during the period. This means that the panel is unbalanced to some degree, which we have illustrated in subsection F.2.

	Measure	Value
1	Number of cases	112663
2	Number of case types	95
3	Number of quarters	46
4	Number of legal changes	345
5	Average absolute change in length	0.09
6	Avg. number of pages (baseline)	1.46
7	Mean error rate	0.40

Table 8: Descriptive numbers

In Table 9 we have highlighted two examples of changes to the paragraphs that constitute the variation we utilize. In the first example, there is some added text and subsections that specify that the municipality must make a decision based on "a holistic assessment" and that it is possible for this assessment to be made by the providing actor. This essentially adds process requirements and another control mechanism of the provided service. In the second example, a reference to another law is added. This makes it possible to use a categorization or decision from another process covered by another law in the decisions. [18]

The two changes above are representative of the variation we exploit. Classifying all 345 changes to these paragraphs over the period, the great majority add cross-references, procedural and documentation requirements, or clearer definitions rather than altering who is eligible for a service. Only one in seven changes adjusts eligibility slightly, and our estimates hold when they are set aside. We therefore interpret the variation as identifying how cases are decided rather than who qualifies.

Measuring complexity and discretion in legal texts

Empirically, scholars have tried different ways to measure legal complexity and discretion.

To measure complexity, we build upon Hurka et al. (2022) and look at (i) a structural measure (length), (ii) a linguistic measure (readability), and (iii) a relational measure (cross-references) of complexity. Furthermore, as our theoretical mechanisms rely on the level of discretion, we also include two measures of different types of delegation that condition municipal discretion.

For the structural dimension of complexity, we use changes in paragraph length as our

[18]The texts are originally in Danish. The translations are our own and were chosen to reflect the changes in the paragraphs. The translations are imperfect and should not be quoted directly.

Example 1: Added text—*The Social Services Act: Substance abuse treatment – Section 101*

Original	Change in 2020
The municipal council must offer treatment to persons with substance abuse problems. Subsection 2. Services under subsection 1 must be initiated no later than 14 days after the inquiry to the municipality. (...)	The municipal council must offer social treatment to persons with a substance abuse problem. **Subsection 2. The municipal council makes a decision on social treatment under subsection 1 on the basis of a holistic assessment of the person's problems and needs** **Subsection 3. The municipal council may request a treatment provider to carry out an assessment of the person's problems and needs, cf. subsection 2.** Subsection 4. Services under subsection 1 must be initiated no later than 14 days after the inquiry to the municipality. (...)

Example 2: Added cross-reference—*The Social Services Act: Choice of placement facility for children – Section 68b*

Original	Change in 2019
The municipal council must make a decision on the choice of the specific placement facility in accordance with the action plan, cf. section 140. At the same time as the placement facility is chosen, a decision must be made regarding the child's or young person's schooling. If the placement facility is located in a municipality other than the child's or young person's municipality of residence, the municipality of residence has a duty to notify the local municipality prior to the placement. (...)	The municipal council must make a decision on the choice of the specific placement facility in accordance with the action plan, cf. section 140 or **the improvement programme, cf. sections 13 and 14 of the Act on the Prevention of Youth Crime**. At the same time as the placement facility is chosen, a decision must be made regarding the child's or young person's schooling. If the placement facility is located in a municipality other than the child's or young person's municipality of residence, the municipality of residence has a duty to notify the local municipality prior to the placement. (...)

Table 9: Examples of changes to the paragraphs

main proxy for added legal complexity because it encompasses several dimensions that all point in the same direction of less discretion. Huber and Shipan (2002) argue that all additional language tends to specify conditions, exceptions, and procedures that narrow implementers' discretion and room to interpret. Laws are typically written densely, so increases in length are unlikely to be redundant filler; rather, they usually reflect added qualifications and constraints. While this measure is coarse at the macro-level, a change in length is more precise within comparable paragraphs. In this sense, paragraph length provides a crude but informative proxy for complexity.

For the linguistic dimension of complexity, we measure how demanding the paragraph is to read and parse, using readability and syntactic complexity indicators. To test the proposed mechanisms of complexity-as-overload vs. complexity-as-clarity, we also use finer-grained measures of complexity. If cognitive strain is the primary source of administrative errors as caseworkers misinterpret laws due to overload, we would expect the error-rate to increase when the law becomes harder to read or syntactically complex, and vice versa. For this, we measure readability through the LIX-score, which is commonly used for the Danish language (Björnsson, 1968; Anderson, 1983). The formula is simple and correlates highly with many other readability tests. Furthermore, we follow Osnabrügge and Vannoni (2025) and measure syntactic complexity through the use of adjectives in legal texts. We also include a measure of contingent clauses (e.g., "if ..., then ..."). This is inspired by Oprea (2020), who argues that complexity stems from the number of "states" and "transitions" that a rule implies, meaning that if a rule has many outcomes (both final and intermediary) and many ways to reach those outcomes, the rule is complex. We mimic this by looking at the number of conditions prior to the basic clause of

a sentence. Using the Python package spaCy, conditions are measured either as a secondary clause starting with the Danish word for "if" or as a secondary clause with the verb preceding the related noun. Empirically, contingencies in laws have been shown to have positive outcomes. In a recent paper, Ash et al. (2025) suggest that more legislation is associated with higher economic growth in U.S. states, and that this effect is driven in particular by contingent clauses. They suggest that contingencies can enhance the completeness of legislation by clarifying the conditions under which rules apply, reducing legal uncertainty and improving predictability.

For the relational dimension of complexity, we capture how interconnected a paragraph is by counting its cross-references to other legal provisions. As noted in the literature on legal complexity, both the length and interconnectedness of laws have increased (Katz et al., 2020; Hinterleitner et al., 2024; Jakobsen and Mortensen, 2015). Following these studies, we also measure the number of references to other paragraphs. Together with the simple measure of length, these two measures are identified by Senninger (2023) as the best predictors of perceived complexity in EU legislative text snippets. A mechanism could be that as the number of cross-references increases, interpretation becomes more coordinated, effectively creating greater legal specificity.

To capture discretion directly, we measure how laws structure bureaucratic discretion through the delegation of authority to municipalities. Following Huber and Shipan (2002), additional statutory detail generally reduces discretion by prescribing conditions and narrowing interpretive room. More recent work focuses on the differences in how delegation is written, as provisions may delegate authority in either a mandatory or a permissive form. A mandatory delegation (using modals such as shall or must) forces bureaucrats to act in a specific way, constraining their choices. By contrast, a permissive delegation (using modals such as may) grants implementing authorities the option to act, leaving more scope for discretionary judgment (Kosti, 2024; Vannoni et al., 2021). Measuring these two types of delegation separately allows us to capture a more fine-grained dimension of complexity and how it structures discretion. To operationalize this, we also use the spaCy package. We classify a sentence as a mandatory delegation if the grammatical subject refers to the caseworker or the municipality as a whole, and the direct predicate is either the modal shall, an active verb, or the phrase "is bound to". A sentence is classified as a permissive delegation if the direct predicate is the modal "can" or "may". We filter out constraints, defined as a delegation but with a negation. Thus, while greater detail through all kinds of additional text may decrease discretion, this approach can help us to distinguish whether complexity clarifies decision-making by constraining discretion (mandatory delegation) or, conversely, creates potential for inconsistency by expanding discretion (permissive delegation).

We have included descriptive information on all of these variables in subsection F.1.

8.4 Research design and empirical strategy

To assess the relationship between legislative complexity and administrative quality, we exploit a panel structure that links complaint outcomes to legal provisions over time. The unit of observation is the individual complaint, nested within a paragraph, municipality, and quarter.

Our outcome is a binary indicator taking the value 1 if the Appeals Board reversed or remanded the municipal decision (an administrative error), and 0 if the decision was upheld.

We estimate the following specification as a linear probability model with two-way fixed effects:

$$\text{Error}_{ijmt} = \alpha + \beta X_{jt} + \mu_m + \gamma_j + \lambda_t + \varepsilon_{ijmt}, \tag{23}$$

where i indexes individual complaints, j indexes paragraph-level case topics, m indexes municipalities, and t indexes quarters. The dependent variable, Error_{ijmt}, equals one if the National Social Appeals Board reverses or remands the municipal decision and zero if the Board upholds it. The main independent variable, X_{jt}, is one of our text-based measures of legal complexity and varies at the case-topic-by-quarter level. For most of the models, we log-transform this variable; a 1% increase in X_{jt} is associated with an approximate $\beta/100$ change in the probability of reversal or remand. We include municipality fixed effects, μ_m, to account for stable differences in local implementation capacity; case-topic fixed effects, γ_j, to absorb time-invariant characteristics of the legal provisions associated with each case topic; and quarter fixed effects, λ_t, to account for common temporal shocks and national trends. Standard errors are clustered at the case-topic level (Wooldridge, 2003).

Each complexity measure is introduced separately. Because our indicators such as length, cross-references, delegation, and readability are highly correlated, including them simultaneously would generate multicollinearity. Estimating models one at a time allows us to identify the direction and magnitude of each dimension.

Conditioning on complaints

Our empirical strategy relies on complaints as quasi-audits of administrative quality. A key validity concern is selection, namely whether the set of citizens who complain changes with legal reforms. If it does, the error rate we observe among complaints might not reflect the underlying error rate among all decisions. Specifically, there are two potential problems with selection.

First, on the extensive margin, if changes to the law also affect the number of people who complain, we might introduce a bias by conditioning on complaints. This requires two conditions: citizens decide to complain partly based on their expected probability of reversal, and true errors raise both the chance of reversal and the propensity to complain. If changes to the legal paragraphs increase the willingness to complain, more marginal cases enter the complaint system, which can reduce the observed reversal rate.

We argue, however, that this is not the case. Theoretically, changes to the law that directly change the pool of relevant citizens would not appear in the data within existing case types but as new case types because, when legislators provide new entitlements, they typically add new paragraphs. Empirically, we can test the empirical implication of an expanded complaint mechanism: legal changes should be associated with changes in complaint volume. We estimate the same fixed-effects specification as the one used in the main analyses, but with the number of complaints as the outcome. Here, we find no meaningful association between our law-change

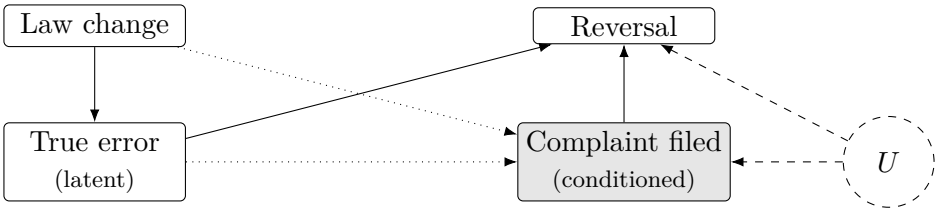


Figure 23: Assumed causal relations between law changes, the true error rate, complaints and reversals. We only observe decisions that citizens complain about, so the analysis conditions on a complaint being filed (grey box). Dotted arrows: paths from law changes and errors to complaints. Dashed arrows: other factors (U) that affect both complaints and reversals.

measures and complaint volume (see subsection F.3). This reduces the plausibility that our main results are mechanically driven by reforms changing the propensity to complain. It might suggest that people complain when they "lose", not only when the municipality errs.

Second, on the intensive margin, if the changes to the law change the composition of the people complaining, we might misattribute a change in the observed error rate to a change in the true error rate. For example, longer laws could make it easier for administratively literate citizens to complain, as they have more rules to support their case, whereas they could decrease the willingness of less administratively literate citizens to complain due to complex rules functioning as a burden. Then the pool of complainers will tilt toward people who file stronger complaints and are more likely to win. That would tend to raise the reversal rate among the remaining complaints, which would bias the estimates upward. The sign of the composition effects is not guaranteed. Although unlikely, it might also be that the citizens who are deterred by complexity are those with the strongest claims, which would make the changes to the composition push the reversal rate down. As always, we cannot fully rule out composition shifts among complainants that leave complaint volume unchanged, but such shifts would have to be strong enough to explain the main findings while generating no detectable change in complaint volume.

Collectively, we propose the following DAG. For this, we argue that since there is no discernible correlation between the number of complaints and the law changes, the dotted edges are unlikely to be quantitatively important in our setting. As such, we argue that the reversal rate serves as a good proxy for the true error rate.

8.5 Results

We begin our analysis of the association between legal complexity and administrative errors by looking at the full linear probability model including all paragraph sections for the entire duration of the panel. We find a statistically significant negative association between paragraph length and the error rate, as reported in Table 10: quarters in which a paragraph is longer show lower error rates than quarters in which the same paragraph is shorter. This could suggest that when the law is made longer, the number of administrative errors declines. A typical change in a paragraph's length (about 9 percent) is associated with an error rate roughly 0.5 percentage points lower. Against a baseline error rate of 40 percent, that is a bit over one percent of the baseline. The coefficient in the first model is -0.054 ($p = 0.016$), so doubling a paragraph's

	Full model	95% trimmed sample	Balanced panel	No time/ municipal FE
log(length)	−0.054*	−0.070**	−0.065**	−0.064**
	(0.022)	(0.026)	(0.023)	(0.023)
Num.Obs.	112 663	108 442	89 294	112 663
R2	0.056	0.058	0.055	0.049
R2 Adj.	0.054	0.056	0.053	0.048
RMSE	0.48	0.48	0.48	0.48
Std. errors	by: CaseTopic	by: CaseTopic	by: CaseTopic	by: CaseTopic
FE: CaseTopic	X	X	X	X
FE: Municipality	X	X	X	
FE: Time	X	X	X	

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 10: Regression results from the linear probability models

length corresponds to an error rate about 3.7 percentage points lower.

In Appendix F.4, a logit model shows the same pattern. To check whether the results are mainly driven by outliers, we restrict our sample to remove the largest 5% of changes to the legal basis. We also include a model with a balanced dataset only including paragraphs that are present throughout all periods in the data set. Furthermore, we include a model without municipal and time fixed effects for comparison. All of these permutations of the model show comparable estimates to the full model.

Across all models, we observe a statistically significant negative relationship between changes in paragraph length and the error rate. In other words, this supports the legal specificity hypothesis: when laws contain more detail, they constrain caseworker discretion and promote more consistent implementation, resulting in fewer overturned complaints. It is hard to reconcile with the complexity-as-red-tape hypothesis, even though our design identifies a within-paragraph association rather than a causal effect.

Unpacking the change in length

As the length of the paragraphs is a coarse measure of the content that is actually in the laws, we further look at the different text-based measures. In subsection F.5 and in Figure 24, we show the results of the linear probability models using the other text-based measures. Because some paragraphs had zero values for mandatory delegation, permissive delegation, conditionals, or cross-references, we add 1 before applying the log-transformation.

If complexity works as red tape, errors should rise with how hard the text is to parse. They do not. The readability coefficient is a precisely estimated null (LIX, −0.002, SE 0.003), and the share of noun-modifiers is likewise indistinguishable from zero (−0.064, SE 0.080). Where the red-tape account predicts its clearest positive signal, we find none.

We find a significant association between the number of cross-references to other paragraphs and laws and the error rate. Across all case topics, the average number of cross-references is 13.6

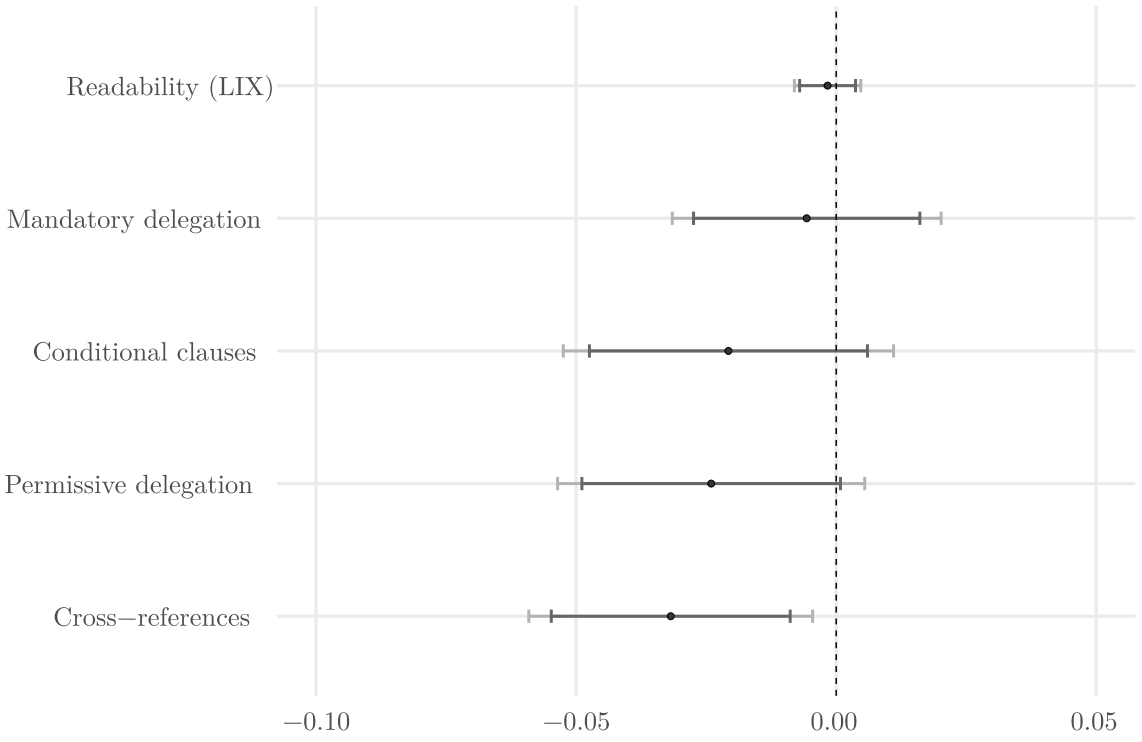


Figure 24: Coefficients for other text-based measures

and the mean change is 25%. With a coefficient of −0.032, the average change in cross-references is associated with a decrease of 0.7pp in the error rate, or about 1.8% of the average error rate. Because the number of cross-references is correlated with paragraph length, it is not certain whether the negative relationship between the error rate and the number of cross-references is direct. Still, this points to cross-references as a candidate textual feature linked to the error rate, and suggests the hypothesis that more references to other laws restrict interpretation, leaving less room for discretion and thereby fewer errors in implementation.

Endogeneity and reactions to errors

A natural concern is that legislators amend the provisions that are already error-prone, so that the negative association reflects politicians reforming error-prone provisions rather than legislative detail-clarifying decisions. This is only a problem if the effect is due to temporary fluctuations and the error rate would have declined anyway, or if attention from politicians has an effect in itself. A cautious test would assess whether high error rates predict subsequent amendments. However, lagged error rates do not predict whether a paragraph is amended in the following quarter (-0.01, $p = 0.72$; see subsection F.6).

However, we do not argue that laws are changed at random. Rather, a probable mechanism could be that politicians change the laws to make them more specific and clearer, which then results in the negative association between length and error rate. This would suggest that politicians are aware of possible mistakes that can be made due to broad discretion, and that this

is something they actually want to ameliorate when amending laws. There is some qualitative evidence suggesting this.

The legislative record is consistent with the reading that these amendments are, in part, responses to administrative error. Across the bills amending these three laws, explicit appeals-and-error vocabulary appears in about two-fifths, against under a fifth of comparable Danish bills, and the National Social Appeals Board is named on the floor in roughly one in five debates (subsection F.7). Legislators voice both positions of our analysis. One minister, introducing a 2024 amendment, said that the rules are so complex that whenever the National Appeals Board looks at them, they find errors. Also, in another session, an MP warned that simplification can remove some legal certainty. [19] Legislators do not argue from the published reversal rates we use as our outcome; the debates, however, often name decisions, audits, and individual cases (Schwalbach et al., 2025).

The consultations also show this. When these laws are amended, the ministry's notes and the stakeholders' responses mention the appeals system and citizens' legal certainty at about twice the rate found in other amending bills generally (subsection F.8). Respondents often argue against "vague" provisions; one respondent described a provision as "an open barn door for unjustified rejections", weakening legal certainty.[20]

Composition changes Regarding whether the changes reflect substantive changes in the eligibility pool, rather than our proposed mechanism, we can also test this. If the association reflected reforms changing who is eligible rather than how cases are decided, removing the eligibility-altering amendments should pull the coefficient toward zero. It does not. Classifying all 345 length-changing amendments and re-estimating without including substantive changes, the estimate is -0.069 (SE 0.032, $p = 0.034$), slightly larger than the baseline -0.054. The identifying variation is therefore not concentrated in amendments that change eligibility (see subsection F.9).

8.6 Conclusion and Discussion

Longer laws are often perceived to be a burden on the people who must apply them, creating difficulties in the implementation and adverse effects on the bureaucracy. We find the opposite in a setting where it can be measured at the level of the individual legal provision: within Danish social-policy paragraphs, more detailed and more cross-referenced law goes together with fewer administrative errors. A red-tape theory predicts that errors rise with length and interconnection; by contrast, we estimate a robust negative association, which fits the theory that longer text leads to clarification. It is hard to square the results with red tape theory.

Our evidence has some natural limits. It is observational and comes from one country and

[19] The quotations are from debates in the Danish Parliament (Folketinget). The English translations are our own and unofficial. Sources: Jens Joel (S), L 108, 2024/25, 17 December 2024; "reglerne er så komplekse, at når Ankestyrelsen kigger på det, finder de fejl i alle sagerne". Pernille Skipper (EL), L 156, 2018/19, 26 February 2019. "...når man afbureaukratiserer i det her tilfælde, tager man også et lille stykke retssikkerhed væk. Altså, man tager en lille ekstra vurdering væk, som er med til at sørge for, at det her er det rigtige at gøre"

[20] The quoted passage is from a *høringssvar* to L 117; the English translation is our own.

one policy domain over a single decade. Also, we cannot identify a clean causal quantity. Of the robust causal-inference designs, the design closest to our setup would be a staggered difference-in-differences design, but it is not well suited to our setting: the treatment is continuous and bidirectional; paragraphs change frequently, so clean pre- and post-periods are scarce, even with binarisation.

However, the findings should theoretically travel well. Denmark has a high bureaucratic capacity, and Fukuyama (2013) argues that the optimal level of discretion rises with capacity: where professional norms and review mechanisms are strong, leaving caseworkers room to judge is less risky than where they are weak. Denmark could therefore be seen as a least-likely case for the claim that constraining discretion helps. If more detailed laws lower error even here, where discretion should be relatively safe, the case for more legal complexity and specificity is plausibly stronger in lower-capacity settings. The broader implication for the simplification agenda is not that more rules are always better, but that cutting law back is not without costs. Legislative detail can help hold decisions up to a common standard.

Collectively, this points in a different direction from much of the literature and government-agency policies. Generally, it is considered to be worse when the law is complex and hard to read. Here, however, it seems that the lengthier the corresponding paragraph is, the more consistent the service delivery. There are obvious caveats. Sometimes a citizen wants flexibility rather than consistency in service provision; in such cases, a "good" law would let the citizen and the caseworker interpret the rules in the citizen's favour, which we cannot measure. From a more legal perspective, however, this suggests that it might be good to have long and detailed laws.











Chapter 9

Increasing Take-up of Social Benefits: A Meta-analysis of Field Experiments

Abstract

Can reducing administrative burdens increase the take-up of social benefits? This meta-analysis reviews 66 field experimental studies reporting 233 treatment effect sizes. Using the administrative burden framework to compare interventions, I reclassify each intervention by the stage at which it measures outcomes (application vs. final receipt) and whether it reduces learning demands by providing information or compliance demands by providing assistance. The results indicate that it is significantly easier to increase application rates than actual take-up rates. On average, estimated treatment effects are about twice as large when outcomes are measured at the application stage as when they are measured at final benefit receipt. The most effective interventions are the ones reducing compliance demands, as these are estimated to increase actual take-up by 7.32 percentage points on average. Interventions reducing learning demands are estimated to increase actual take-up by 3.17 percentage points on average. These findings consolidate the field experimental evidence on how to improve take-up rates and highlight the need for further research on application stages, treatment compliance and variation across welfare regimes.

9.1 Introduction

Many eligible individuals fail to receive social benefits due to lack of information, complexities, and inefficiencies in bureaucratic processes (Ko and Moffitt, 2024; Currie, 2004; Oorschot, 1991). This has been an issue for a long time and in various countries with different welfare regimes (Bargain et al., 2012; Harnisch, 2019; Ko and Moffitt, 2024). This failure not only compromises legal entitlements and exacerbates inequalities (Deshpande and Mueller-Smith, 2022; Aizer et al., 2016; Hoynes et al., 2015; Hoynes et al., 2016), but also undermines the intended impact of social programs designed to support vulnerable populations. Understanding both the barriers to accessing government services and the methods for helping citizens overcome them is thus vital for creating equitable and effective public programs.

A large body of research has used the framework of administrative burdens to study take-up of social benefits (Herd and Moynihan, 2018). The framework conceptually aligns research on individuals' experiences of policy implementation as onerous under a single framework, highlighting the commonality in the challenges across different administrations. Within this, Bækgaard and Tankink (2022) distinguish between state actions and citizens' experiences. The state imposes learning demands (information that citizens must understand) and compliance demands (processes that they must complete), which in turn impose costs on citizens. When people cannot meet these demands, they may forgo benefits even though they are eligible. The framework also encapsulates the early work in economics trying to model different citizen specific, policy specific and institutionally specific characteristics that could theoretically hinder take-up, e.g. welfare stigma (Moffitt, 1983), imperfect information (Bhargava and Manoli, 2015), complexity (Kleven and Kopczuk, 2011), and transaction costs (Currie, 2004). Various studies have tested different interventions to increase take-up across a wide range of programs and countries, and the evidence is mixed (Daigneault et al., 2025). For example, Guyton et al. (2017), Bhargava and Manoli (2015), and Beecroft (2012) find that simple outreach messages increase the take-up of Earned Income Tax Credit (EITC) in the United States, whereas these findings are not replicated in a large, similar study by Linos et al. (2022) or Cranor et al. (2019) for the same benefit.

The aim of this paper is to consolidate this large body of evidence on take-up from various heterogeneous contexts and across fields in the social and behavioral sciences. This paper thus provides a reinterpretation and meta-analysis of field experimental studies aiming to figure out how to increase take-up of social benefits. Each intervention changes what citizens encounter so I classify them by whether they reduce learning demands by supplying the necessary information or reduce compliance demands by offering assistance. Furthermore, I distinguish between studies that measure citizens' behavior early in the process, during the application phase, and studies that measure behavior at the end of the process, when people have incurred more burdens and eligibility checks to either receive the benefit or not. There is some value in the total average effect, as individual studies can lack both precision and external validity when they are conducted in only one setting and sometimes have a relatively low number of observations per treatment arm (Slough and Tyson, 2023).[21] However, the main purpose of this paper is to study some

[21]The overall estimated average effect across studies is 5.97 percentage points.

of the shared heterogeneity explaining the differences between studies, as this can highlight gaps and guide how to most effectively alleviate missing take-up. Explicitly focusing on study heterogeneity is a common strategy of meta-analyses (DellaVigna and Linos, 2022; Galos and Coppock, 2023; Backhaus and Vogel, 2022). By comparing studies across multiple benefits and institutional contexts, this paper both highlights their external validity and identifies key factors to target in order to increase take-up of social benefits.

I find that studies that measure outcomes at the application phase find significantly larger effects than those measuring actual take-up. Studies measuring application have an estimated average difference between the control and treatment group of 9.14 percentage points compared to 4.42 percentage points in the studies measuring actual take-up. Measured using risk ratios, this corresponds to an increase of 27.7% for the studies measuring actual take-up and 47.4% for the application studies. The large estimated difference between the two groups of studies suggests that part of the problem of take-up lies after the initial application process in either subsequent steps or eligibility checks. Furthermore, I find that the estimated average effect of interventions reducing compliance demands is 7.32 percentage points, which is significantly larger than the effect of 3.17 percentage points for the interventions reducing learning demands. Collectively, this highlights the importance of trying to help people overcome the compliance demands of adhering to administrative rules and requirements, such as completing forms and providing documentation for a social benefit. Even so, there is still a significant positive effect of providing information and raising awareness. Finally, I discuss different forms of publication bias. Attempting to correct for them reduces the estimated effects overall; however, the effects remain statistically significant.

9.2 The study of take-up

Social benefits, designed to alleviate economic hardship and promote social equity, are found across modern systems of social provision. Whether in the form of food assistance, tax credits or health insurance subsidies, social benefits across the globe operate under similar principles and face comparable challenges. However, another characteristic is also comparable worldwide, namely, the persistent issue of non-take-up (Ko and Moffitt, 2024). Social assistance programs look very different on the surface. Some, such as the EITC in the United States, are claimed simply by completing a short schedule form on an existing tax return. Others, for example the Supplemental Nutrition Assistance Program (SNAP) or Temporary Assistance for Needy Families (TANF), require a separate, multi-step application, documentary proof of identity, income, employment, and often in-person interviews. Although the extent and form of these requirements differ, most programs follow two common stages: individuals first learn about the benefit and submit an application; and, if they pass eligibility checks, they receive the benefit.

Daigneault (2024) highlights the importance of acknowledging this implicit "stage logic" in the study of resident-state interactions, as the process of receiving a benefit requires overcoming burdens at multiple stages. In other words, the application process accumulates administrative burdens over multiple stages. Studies that try to raise social-benefit take-up measure different

points in that process; some studies stop at application initiation, while others measure actual benefit receipt. Studies measuring actual take-up necessarily also capture whether people have applied for the benefit. Consequently, effect sizes for actual take-up should generally be smaller than those for applications alone. However, the magnitude of this difference is theoretically interesting. Large differences in the ability to raise application rates without raising actual take-up could be explained by two factors: more ineligible applications or drop-off after application. Interventions may prompt people who do not qualify for the benefit to apply, or they may get eligible people to apply who then fail at later administrative stages, ultimately hindering actual take-up. As such, the difference reflects part of the gap between the ability to change people's behavior and the ability to help people actually overcome administrative burdens. Acknowledging this difference is important when drawing conclusions from interventions on what works to increase the take-up of social benefits.

Interventions and administrative burdens

Interventions in this field all try to do the same thing: increase take-up. However, because they come from different fields, they use different terminology for the underlying theoretical mechanisms. In a narrative review, Daigneault et al. (2025) describes and compares many of these in detail. There is necessarily heterogeneity between singular studies that focus on different ways to combat low take-up; however, using the administrative burdens framework (Herd and Moynihan, 2018), one can understand these barriers as similar mechanisms. In this framework, citizens experience the state's actions in different ways, making the interaction onerous. Typically, the literature focuses on three main categories: learning costs, which involve gathering information about programs and eligibility; psychological costs, such as stress, stigma, or loss of autonomy; and compliance costs, which refer to the effort required to navigate administrative rules and complete necessary paperwork or documentation. These categories are broad and suggest that costs can be incurred at multiple stages, including before, during, and after interacting with the government. In terms of social benefits, if citizens cannot overcome these costs, they will not receive the benefit. This provides a practical metric for scholars studying administrative burdens, as the findings from take-up studies are also relevant for a wide range of other outcomes (Halling and Bækgaard, 2024), e.g. democratic participation (Selin, 2019), immigration (Heinrich, 2018), digitalization (Peeters and Widlak, 2018; Peeters and Widlak, 2023) and healthcare (Kyle and Frakt, 2021).

The different costs can arise from different parts of the process and are not mutually exclusive. Bækgaard and Tankink (2022) distinguish what the citizen experiences from what the state does, highlighting that e.g. reducing compliance demands can both decrease the psychological costs of stress and the compliance cost directly associated with the demands. The citizen encounters demands and experiences costs. While many studies published after the introduction of the framework do cite the administrative burden framework and theorize their mechanisms through this, there is still a large body of research on missing take-up that either predates it or simply does not cite it. Mapping all take-up studies onto the same burden categories could consolidate the literature and provide a unified theoretical interpretation.

Field experiments have been the most robust method used extensively for studying the empirics of take-up of social benefits. These experiments are particularly valuable in demonstrating how take-up can be increased through various types of interventions and, especially for academic purposes, in that they create exogenous variations in real-world environments. This allows researchers to study differences in perceptions and experiences of real-world phenomena that could not otherwise be directly changed (Haaland et al., 2023). As such, looking at which treatments work provides guidance on how to increase take-up and also helps identify which government characteristics citizens find especially difficult to deal with. The individual studies that reference the administrative burden framework typically describe their treatments as targeting specific costs that citizens experience. For example, studies that frame information about a social benefit in a way that tries to de-stigmatize eligible people are targeting psychological costs. However, looking at the commonalities between the studies, I reinterpret the field experiments as introducing variation in what the residents *encounter* rather than what they actually experience. By introducing variations in the application process through information dissemination, scholars are actually measuring the effects of lowering *learning demands*, as residents now have more readily available and appropriately simplified information about a benefit for which they are likely eligible. Similarly, when studies provide assistance to people, the experiments isolate the effects of lowering *compliance demands*, as residents now have to do less work in the application process by themselves. Reinterpreting interventions as changes in what residents actually encounter during the application process makes them more comparable across these different contexts.

Relying primarily on field-experimental treatment effects makes sense for a meta-analysis for several reasons. Field experiments provide a simple, causal measure: the increase in percentage-point take-up compared to a control group. This standardization makes it easier to compare different studies, as they all report effects on the same scale from 0 to 1. This advantage of field experiments is also exploited in (DellaVigna and Linos, 2022). Prioritizing high-quality, causal field experiments helps to avoid the pitfalls of averaging mixed-quality studies, highlighted by Simonsohn et al. (2022) as a problem for meta-analyses. This selective, design-focused approach ensures that the conclusions are not obscured by weaker correlations but instead grounded in robust evidence. Whereas quasi-experiments relying on naturally occurring variations offer similarly credible causal estimates, and provide indirect insights into counterfactual outcomes, the treatments are often clustered in design and more time-sensitive, making their effects less comparable across settings than those observed in field experiments. Although field experiments are not without their faults, they are the best evidence available on how administrative burdens can be relieved to increase take-up.

9.3 Methods

To find studies for the meta-analysis, I relied on several sources. I searched Web of Science using a search string of words commonly associated with take-up studies (see Appendix G.3). I used this search string again using SSRN and OSF in order to avoid publication bias and include working papers. To find studies not made for academic journals, I searched websites for various

nudge units using an OECD list (Hubble and Varazzani, 2023) and the studies included in DellaVigna and Linos (2022). Furthermore, I reviewed forward citations to early articles in the field, such as Currie (2004), Schanzenbach (2009), Bhargava and Manoli (2015), and Bettinger et al. (2012). In Appendix G.1, the literature search and number of exclusions are summarized.

I excluded studies for multiple reasons. As I want to compare effects from different contexts, but with the same type of intervention, it is important that the outcomes across studies are aligned. I exclude studies that are not actually social benefits. Mostly, I only include studies that are essentially "free money"; however, I do also include some highly subsidized services such as health insurance. The strict selection criteria help to rule out lack of demand as a reason for low take-up. I do not include uptake of job training programs, savings accounts, breast cancer screenings, or medical visits even though doing so would have greatly increased the number of eligible studies. This harmonization minimizes the risk of combining studies with inherently different outcomes or treatment comparisons, thus enhancing external validity while at the same time ensuring variation across interventions.

I have excluded studies, some seminal, wherein there is no clear control group, e.g.: Bhargava and Manoli (2015), Alatas et al. (2016), Aizer (2007), and Lerman et al. (2017). This increases measurement harmonization, as I compare the simple percentage-point difference between the control group and the treatment group across settings. Similarly, I have not included quasi-experiments even though they estimate causal effects, because they typically evaluate a program change rather than a change in citizens' perceptions, levels of information, or engagement with the programs, making them less comparable (Shepard and Wagner, 2022; Leroy, 2024; Armour, 2018; Finkelstein et al., 2019; Homonoff and Somerville, 2021). Lastly, I also only measure binary outcomes, i.e. whether a person does or does not receive a benefit, excluding studies that measure continuous levels of uptake such as the amount of money received. These studies are all very valuable for answering questions about the role of administrative burdens in hindering take-up; however, including them in a meta-analysis is not possible.

As some studies test more than one treatment, I end up with more effect sizes than studies. In total, the final data set includes 233 treatment effect sizes within 66 individual studies that all try to heighten take-up in one way or another.

The dissertation update adds 46 effect sizes from 15 studies identified and adjudicated after the original search; these additions are included in every reported estimate and generated output below.

Coding of application stage and administrative burdens

In order to do any meta-analysis on outcomes and interventions that are not exactly the same, it is necessary to simplify study characteristics into comparable categories with common metrics. As such, I code the different interventions according to whether they increase application rates or take-up rates, and whether they lower learning or compliance demands.

To determine whether studies measured application or actual take-up, I rely on the studies' description of whether they had access to data proving that people actually received these benefits. For example, for studies using the EITC, I look at whether the authors had access to

data on payout or whether they relied on the participants themselves stating a claim. Within the two categories of application and actual take-up, the studies also differ. Some studies, such as De La Rosa et al. (2021), measure the initiation of an application rather than the actual submission, as most other studies do. For some benefits, such as SNAP, the application stage is followed by later stages of interviews and further eligibility checks, whereas in e.g. the rental assistance benefit studied in Lasky-Fink and Linos (2024), sending an application is the final step before actually receiving the benefit, provided the resident meets the eligibility check. A clearer case is the studies that examine the take-up of FAFSA, where some studies measure whether a student received a grant, whereas others merely look at the filing. Within the application category, I treat the different types equally since I study how interventions change residents' behavior, but this difference is important when drawing conclusions.

Few studies test the exact same benefit. To determine whether the authors are reducing learning or compliance demands, I look at whether the intervention makes the application process less onerous and simpler or merely provides information. Here, I do not distinguish between whether the studies frame information differently or include further nudges, e.g. to reduce stigma. However, there are some grey areas. For example, some studies provide information about access to assistance such as Volunteer Income Tax Assistance (VITA) (Goldin et al., 2022; Linos et al., 2022). Other studies provide assistance via telephone or at a physical location. For the main analysis, I treat information about assistance as decreasing learning demands, as there is no guarantee that it actually makes the application process simpler, unlike the other studies providing direct assistance. Although assistance interventions often include informational elements, I code learning-demand and compliance-demand interventions as mutually exclusive.

Estimation

In the meta-regressions, each observation is the simple difference in take-up between a treatment arm and its control group. This is the intention-to-treat (ITT) effect, $\hat{\theta}_{ij}$, expressed in percentage points, for each treatment arm i in study j. This estimate is also the main effect most studies report.[22] If a study does not report this information, I estimate it. For example, if not provided I infer the exact number of participants in each treatment arm from the totals, or I convert a reported complier average treatment effect (CATE) into an ITT by multiplying by the compliance rate. I furthermore ignore subgroup analyses, added control-variable regressions, and event-study specifications unless they are presented as separate experiments with their own control groups to keep only one comparable estimate per treatment group. As I measure the percentage difference between control and treatment group, I include studies that measure on loss of benefit also. I rely on the authors' aggregation if they do not report the effects for each treatment group. Where studies report more than one treatment effect, I cluster standard errors on the study-level.

Using a standard three-level meta-analysis model, I estimate the average effect using the

[22]For the main models I use the absolute percentage-point differences rather than risk ratios because that scale is easier to interpret and avoids overstating effects in studies where control-group take-up is low. Appendix G.6 in the appendix plots each effect size against the control-group take-up rate. Risk-ratio estimates become strongly right-skewed as the baseline rate falls, while percentage-point differences remain roughly symmetric. However, I still include the results using risk ratio in the appendix

following equation:

$$\hat{\theta}_{ij} = \mu + \beta X_{ij} + \zeta_{ij}^{(2)} + \zeta_{j}^{(3)} + \epsilon_{ij} \tag{24}$$

$\hat{\theta}_{ij}$ represents the estimated effect size for study j and within-study observation i. X_{ij} is a binary indicator equal to 1 if the outcome is actual take-up and 0 if it is application. The coefficient β therefore captures the average difference between application and actual take-up outcomes. The model incorporates two heterogeneity components: $\zeta_{ij}^{(2)}$, representing the variation in true effect sizes within studies, and $\zeta_{j}^{(3)}$, representing the variation between studies. The model is estimated with restricted maximum likelihood (REML) using the *metafor* package, which weights each observation by the inverse of its sampling variance, so that observations based on larger samples receive more weight.

I keep all the effect sizes, each treatment-versus-control contrast, in one three-level meta-regression. The estimates are comparable as they measure the same type of effect (ITT in percentage points). Retaining every arm, rather than averaging within studies, keeps valuable variation and sharpens the tests of moderators such as application versus actual take-up and subsequently whether the intervention reduced learning demands or compliance demands.

The hierarchical specification (treatment arms nested within studies) lets the random-effects terms account for the correlation that arises when several arms share a control group or other design features. Cluster-robust CR2-standard errors then correct for any remaining within-study dependence. Pooling the data in this way therefore uses all available information and provides the most precise picture of how take-up interventions work across different contexts.

It should be noted that this meta-analysis should be interpreted as a cross-sectional comparison of studies (Slough and Tyson, 2023). This comes with some limitations. Relying only on field experiments might skew the results as they often rely on external collaborators who also have an interest in increasing take-up of the benefits. This is an external-validity concern of the causal claims of field experimental studies as there exists some degree of selection into the literature. Although pooling studies in a meta-analysis might reduce this problem, systematic differences across settings can bias the combined estimates. Furthermore, as a cross-sectional model, the estimates cannot be interpreted causally even though the individual studies estimate a causal effect. They are weighted aggregates of study-level estimates drawn from different populations, interventions, and settings, which the models try to control for. As with all cross-sectional models, it is not possible to control for everything. If the observed study features correlate with unmeasured factors, the random-effects variance absorbs some of that extra heterogeneity and inflates the standard errors, but any remaining correlation can still bias the pooled effect. As the analysis is comparative in nature, it thus does not provide a more credible estimate of the studies' actual effects, but rather a summary across the various characteristics of the interventions. Each study yields an internally valid estimate of each of its own interventions and much of the variation in effect sizes comes from contextual heterogeneity, as captured in the model above. However, the studies are still comparable. Following Slough and Tyson (2023), the studies are harmonized as they measure the same type of outcome (benefit take-up) and target-equivalent since they all use the same treatment-vs.-control contrast. Thus, because the studies share some commonalities, the meta-regression can help answer how the effect sizes vary

Description	Actual take-up	Application
Number of studies	51	29
Number of effect sizes	148	85
Median of individual effect sizes (in pp)	1.4	2.7
75th percentile of individual effect sizes (in pp)	4.75	9.1
25th percentile of individual effect sizes (in pp)	0.27	0.52
Mean treatment group size across studies	165975.8	32651
Median treatment group size across studies	2383.5	10740

Table 11: Descriptive statistics by outcome type

Statistic	Compliance	Learning
Total number of studies	25	53
Total number of effect sizes	45	188
Actual take-up: number of studies	21	41
Application: number of studies	9	24
Actual take-up: number of effect sizes	32	116
Application: number of effect sizes	13	72

Table 12: Number of studies and effect sizes categorized by whether the interventions are reducing learning demands or compliance demands

across these common features such as operationalization of the outcome or intervention type.

9.4 Results

Descriptive statistics

As seen in Table 11, most studies measure actual take-up within the definition provided above. Fourteen studies provide estimates of both application and actual take-up. The application effect sizes typically come from larger studies with more people in the treatment groups. Looking at the different percentiles of the effect-size distribution suggests that the effect sizes of the studies measuring application are larger than those measuring actual take-up.

The number of studies varies across different characteristics. As shown in Table 12, the number of studies reducing learning demands is lower than that for studies reducing compliance demands. There are especially few of these within the subgroup of studies measuring application, making it more difficult to provide precise estimates of the difference within this subgroup.

Similarly, the field of take-up of social benefits is more studied in an American context. This might be due to the American welfare system being generally less universal than those of many other countries; however, the imbalance is striking given the global nature of the issue (Ko and Moffitt, 2024). In Appendix G.2 the studies are listed by individual US states. While California is overrepresented there is a broad selection of states in this literature.

There is a wide variety of benefit types included in the dataset. Most studies look at tax credits such as EITC, all of which are from the US. Subsidized health insurance studies are the most internationally diverse type of benefit, whereas student benefits (such as those obtained through FAFSA) and food stamps (SNAP) are also studied only in the US. All the benefits require an application and are means-tested to some degree.

Country	Effect sizes	Studies
USA	157	45
Australia	12	1
China	12	3
India	12	3
France	8	3
Sweden	8	1
Belgium	6	1
Indonesia	4	1
Argentina	3	1
Ghana	3	1
Germany	2	1
Uruguay	2	1
Colombia	1	1
Finland	1	1
Philippine	1	1
Vietnam	1	1

Table 13: Number of studies and effect sizes by country

Benefit type	Effect sizes	Studies
Tax credit	52	10
Subsidized health insurance	43	15
Student benefit	41	9
Food stamps	23	10
Pensioner benefit	20	4
Basic income	19	4
Child Benefit	11	4
Housing Benefit	11	3
Mix of multiple services	8	5
Free schooling	4	1
Loan-forgiveness	1	1

Table 14: Number of studies and effect sizes by type of benefit

Estimated average effects

Looking at both studies with application and studies with actual take-up as outcome, running the meta-analysis regression suggests that there is an estimated (inverse-variance weighted) average effect of 5.97 percentage points across all treatments. This is unsurprisingly highly significant as most of the studies have a positive effect (see Appendix G.5). Looking at the difference between studies by outcome type in Table 15 and Figure 25, studies that measure application have an estimated average effect of 9.14 percentage points, which is significantly larger than the estimated average effect of 4.42 percentage points from the studies that measure actual take-up. The difference reflects the findings from the studies that examine both outcomes, such as Linos et al. (2024). These effect sizes are both larger than the median of each subset, as the estimated effect lies at approximately the 75th percentile for studies measuring application and around the 74th percentile for studies measuring actual take-up. Measured as risk ratios in Appendix G.5, this corresponds to an increase of 27.7% compared to the control group for the studies measuring on actual take-up and 47.4% for the application studies. Looking at the forest plot Figure 26, it is evident that there is a large level of heterogeneity between effect sizes across studies, with the subset of studies measuring actual take-up having the most outliers. This could mean that the estimates are being inflated by the outliers. Many studies in the sample report low, non-significant treatment effects; conversely, a few studies report substantially larger effects, such as Milap and Sarin (2023), who study an intervention where they provided assistance to households to get free schooling in India, or Matikka and Paukkeri (2022), who study informational treatments leading to increases in take-up exceeding 30 percentage points in the case of a low-income pensioners' benefit (see Table 42). Table 42 in Appendix G.1 gives more information on the individual points in the forest plot.

To examine the influence of outliers, I include some robustness checks in Table 15 and Figure 25. Excluding the studies with the top 10% largest Cook's D, however, does not change the difference in effect sizes between the two types of outcome, although the overall estimated

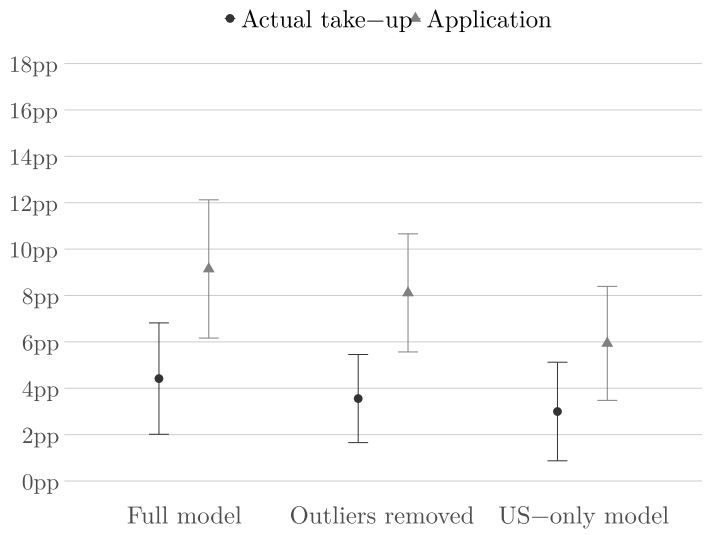


Figure 25: Predicted Estimates by Outcome Type

Note: These values are the predicted estimates from the models in Table 15.

	Full interaction model	US-only model	Outliers removed
Intercept (Actual take-up)	0.044***	0.030**	0.036***
	(0.012)	(0.011)	(0.010)
Application	0.047*	0.029+	0.046*
	(0.015)	(0.014)	(0.013)
Num.Obs.	233	157	209
SE type	CR2	CR2	CR2
Cluster level	Study	Study	Study

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 15: Regression table: Difference between studies measuring on application vs. actual take-up

effect decreases. As the findings from the studies outside the United States are generally larger, I run a model with the subset of studies only from the United States. The directional difference between the two types of studies is the same as before; however, the difference is marginally insignificant at the 10%-level. All models still suggest that there is a significant positive effect of increasing take-up even when measuring final actual take-up.

Although there is a lot of heterogeneity decreasing the statistical precision, the difference between the estimates from the two subsets is quite large across model specifications. These models suggest that the estimated average effect of increasing application rather than actual take-up is around twice as large. This will not be the case in every study and with every measure of application, since some of these are closer to actual take-up than others. Even though the classification of application and actual take-up is a simplification in trying to measure across different studies and contexts, the results do suggest that it is significantly easier to get people to take the first step in taking up a social benefit rather than actually helping them receive it. Yet, it is hard to disentangle whether this difference stems from treatments simply getting ineligible

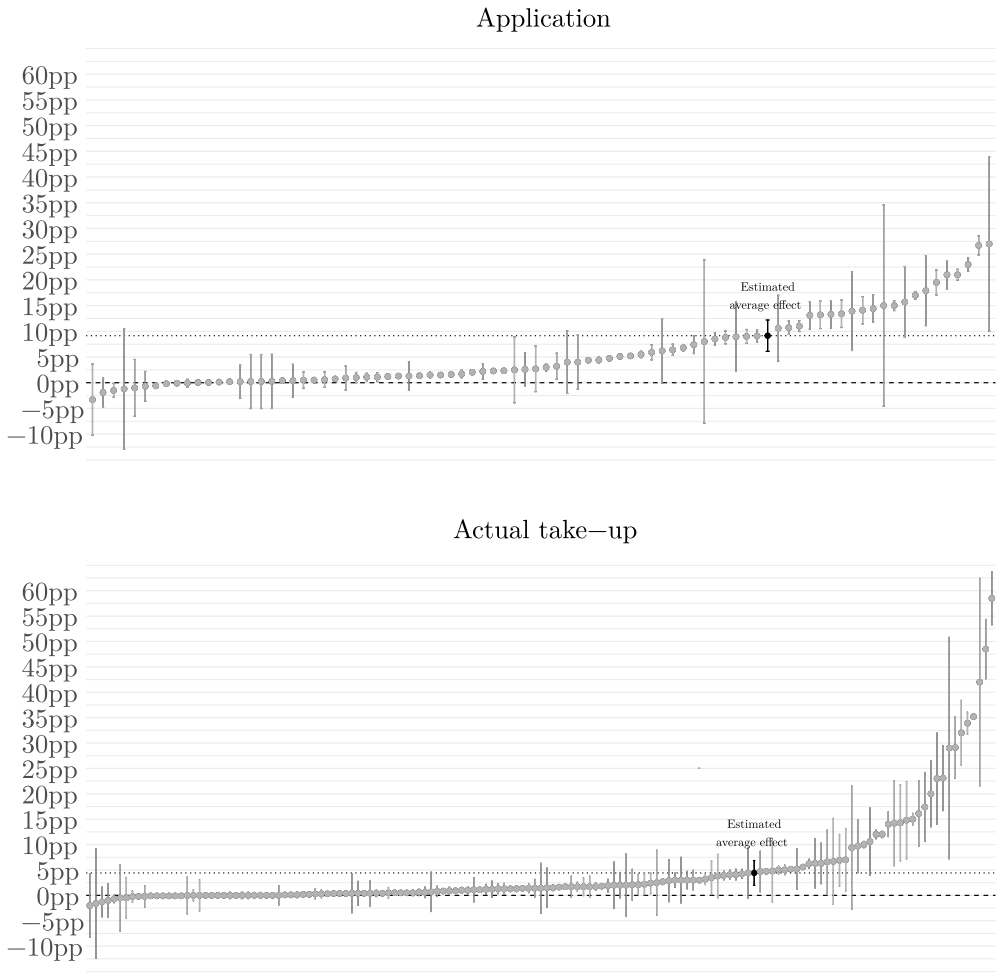


Figure 26: Forest-plot of effect sizes

people to apply or from people failing to complete further steps on their way to receiving a benefit. In Appendix G.6, I do an exploratory analysis by including a variable indicating whether the benefit application has extra steps after submission. Although underpowered, the estimates directionally suggest that there are larger effect sizes in studies with extra steps. Some of the studies that provide a measure of both the application and the take-up rate (such as Van Gestel et al. (2023) and Engström et al. (2019)) show a significant treatment-control difference in the conversion of applications into take-up, suggesting that some of the variation at least stems from ineligible people applying at a greater rate in the treatment groups. This highlights the importance of precise eligibility estimates prior to rolling out treatments.

Decreasing learning demands or decreasing compliance demands

Next, I examine differences in estimated effects across intervention types. In Table 16, the intercept represents the estimated average effect of learning-demand interventions measuring actual take-up. For all specifications of the models, this is estimated to be significantly different from zero, suggesting that on average, across contexts and benefits, there is a positive effect of

	Across outcomes	Full interaction model	US-only model
Intercept (Actual take-up)	0.045***	0.032**	0.017+
	(0.010)	(0.012)	(0.009)
Application		0.043*	0.021*
		(0.017)	(0.007)
Compliance	0.056*	0.041*	0.044+
	(0.019)	(0.014)	(0.018)
Application × Compliance		0.035	
		(0.057)	
Num.Obs.	233	233	157
SE type	CR2	CR2	CR2
Cluster level	Study	Study	Study

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 16: Regression table: Reducing learning demands or reducing compliance demands

outreach in increasing take-up. Interventions lowering compliance demands have substantially larger effects than those lowering learning demands. Across all measured outcomes, studies that try to reduce compliance demands have 5.59-percentage-point larger estimated effects. Measured as an interaction, the meta-regression results suggest that for studies increasing actual take-up, there is a 4.15-percentage-point larger effect in the compliance-demand studies. Measured as risk ratios in Appendix G.5, the estimates suggest an average effect of increasing take-up by 21.7% compared to the control group when reducing learning demands and 46.5% when reducing compliance demands in studies measuring on actual take-up.

For the US-only model, the results are directionally the same, although the effect sizes are smaller and both the intercept and the estimate for reducing compliance demands are only significant at the 10%-level. As seen in the descriptive statistics in Table 12, there are not a lot of studies measuring application that provide assistance and reduce the compliance demands for the treated people. Consequently, I cannot precisely estimate this subgroup difference. However, the positive point estimates cautiously suggest even larger effects of reducing compliance-demands at the application stage.

Defining what counts as a reduction in compliance demands is important, but not straightforward. Some studies are trying to provide assistance to residents through letters stating where they can contact someone, if they are in need of assistance. However, for some people this might be perceived as yet another step to comply with before receiving a social benefit. In the main regressions in (Table 16), I have included telephonic guidance, although this might also not actually help citizens as they themselves are still the ones completing files and filling out forms. I test different conceptualizations of what actually constitutes a reduction in compliance demands in an intervention in Table 17. The difference in the results could suggest that the results are driven mainly by the interventions using in-person assistance, as the coefficient for reducing compliance demands is larger in this model. Similarly, the coefficient is smaller and insignificant in the model using the broad conceptualization by including interventions that simply send invites to receive assistance. Collectively, this shows that the effects of increasing take-up are

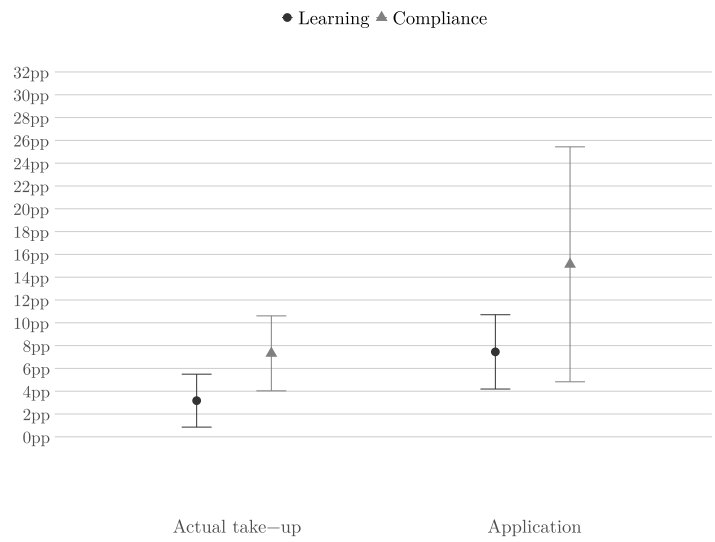


Figure 27: Predicted Estimates by intervention types: Reducing learning demands or reducing compliance demands

Note: These values are the predicted estimates from the full interaction model in Table 16.

larger in studies that to a larger extent guarantee that people receive the assistance they need.

Publication Bias

In the analysis above, I have included both working papers and studies that were not intended for publication in academic journals, in order to mitigate sources of publication bias. These steps, however, are not perfect and only address bias arising from the likelihood of journal publication, not the possibility that RCTs have been conducted but never written up as manuscripts—also known as the "file drawer" problem (Rosenthal, 1979; Franco et al., 2014). Since the analysis is limited to RCTs, p-hacking may be less of a concern: Brodeur et al. (2020) estimate that only 1.5% of RCT test statistics are likely p-hacked or selectively reported. The more relevant issue is the potential existence of unpublished RCTs with null or near-zero effects that are excluded from the dataset. If such studies are systematically missing, the estimated effects will be inflated regardless of the heterogeneity modeled above.

Because publication bias cannot be observed directly, it must be inferred under the assumption that sample size and effect size are uncorrelated. I use two methods to analyze publication bias: (1) visual inspection of funnel plots, and (2) the method by Andrews and Kasy, 2019. Firstly, ignoring the heterogeneity in the studies, the funnel plot in Appendix G.4 is indeed asymmetrical, which is indicative of publication bias. The funnel plot also shows that there are many precise null findings, however. It is similarly visible that both precise and imprecise outliers may be skewing the estimates. This is an inherent limitation of the analysis, as effect size estimates are bounded below near zero, resulting in a distribution that is inherently right-skewed even without publication bias. This bias can be somewhat corrected by excluding outliers in Table 15. This does not, however, mitigate the file drawer problem as the analysis might be missing studies with effect sizes closer to zero. I use the method of Andrews and Kasy, 2019 to test for this,

	Main model	Broad: Invites included	Narrow: Only in-person
Intercept	0.032**	0.032*	0.032**
	(0.012)	(0.013)	(0.011)
Application	0.043*	0.045*	0.046*
	(0.017)	(0.017)	(0.017)
Compliance	0.041*	0.034+	0.067**
	(0.014)	(0.017)	(0.018)
Application × Compliance	0.035	0.010	0.054
	(0.057)	(0.043)	(0.072)
Num.Obs.	233	233	233
SE type	CR2	CR2	CR2
Cluster level	Study	Study	Study

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 17: Regression table: testing conceptualizations of reducing compliance demands

as it models the probability of publication as a step function of statistical significance (with a threshold at $z = 1.96$) and reweighs the observed estimates accordingly. There will be some imprecision in this, since I have not corrected standard errors for the additional controls and clustering that the individual studies might have reported. Across all studies, the models in Appendix G.4 suggest that results that are insignificant at the 5%-level are around 12.1% (SE = 0.069) as likely to be reported as significant results. Correcting for this selection gives an estimated mean of the distribution of true effect sizes of 2.1 percentage points (SE = 0.86 percentage points). By only looking at the subset of studies that are measuring actual take-up, this effect size drops to 1.6 percentage points (SE = 0.64 percentage points). This replicates the findings from DellaVigna and Linos (2022), who study a broader range of behavioral nudges. With the updated data, the application-only optimization is numerically unstable and therefore does not support a credible corrected subgroup mean.

Nevertheless, these estimates might be overly cautious and attenuated. While the data do show a negative correlation between precision and effect size, this might not be solely due to publication bias, but instead to the heterogeneity in the studies. As most researchers give some thought to statistical power before conducting a study, it might be that if they are testing more intense and presumably more expensive interventions (such as the ones targeting compliance demands), they will not use a sample size larger than necessary for estimating a somewhat precise effect. As such, the assumption that the sample size and effect size are uncorrelated may not be valid in meta-analyses modeling substantial heterogeneity. Furthermore, following DellaVigna and Linos (2022), the academically targeted estimates are not significantly larger than the ones from Nudge Units that are less likely to be concerned with statistical significance, reported in Table 18.

Another potential source of bias in the estimated effects is that the true impact of interventions to increase take-up may have declined over time. This pattern is often referred to as "the decline effect" (Schooler, 2011) or "voltage drop" (List, 2022), and the phenomenon is common across

	Nudge Unit model	Nudge Unit interaction	Decline effect	Decline effect (US-only)
Intercept	0.059***	0.048**	0.153**	0.078+
	(0.013)	(0.014)	(0.046)	(0.038)
Application		0.036*	0.087	0.086
		(0.013)	(0.045)	(0.054)
Nudge Unit	0.002	−0.019		
	(0.021)	(0.027)		
Years since 2006 (earliest intervention)			−0.008*	−0.003
			(0.003)	(0.003)
Application × Years since 2006			−0.004	−0.005
			(0.003)	(0.004)
Application × Nudge Unit		0.037		
		(0.043)		
Num.Obs.	233	233	80	53
SE type	CR2	CR2	CR2	CR2
Cluster level	Study	Study	Study	Study

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 18: Regression table: test of different sources of publication bias

different fields when trying to scale up interventions. A common source of this drop is that interventions might work differently across different populations. This highlights both an advantage and a disadvantage for meta-analytic studies, as the primary aim is to look across and model various types of heterogeneity, but this might bias the overall estimates leading to false inferences. The pattern could be explained by publication bias where early studies reported large, attention-grabbing effects in most-likely contexts that spurred replications with smaller impacts. Alternatively, it could result from changes in the broader context such as improvements in governments' ability to identify eligible residents or efforts to reduce administrative burdens through more accessible application processes. In Table 18, I take only the largest effect size from each study of each measure and include a variable for the year the study was conducted and find a significant negative association between year of intervention and effect size. The finding is directionally the same in the US-only model, however it is statistically insignificant due to the loss of power. While this finding is exploratory and sensitive to inclusion and exclusion of studies, this does suggest that the overall estimates are being amplified by early interventions with a high impact.

The overall estimates are heterogeneous, and another source of this heterogeneity is the time of the intervention. Taken together, however, while there might be biases in the meta-analytical models, attempting to correct for publication bias still yields positive, statistically significant estimates for interventions that aim to raise take-up of social benefits.

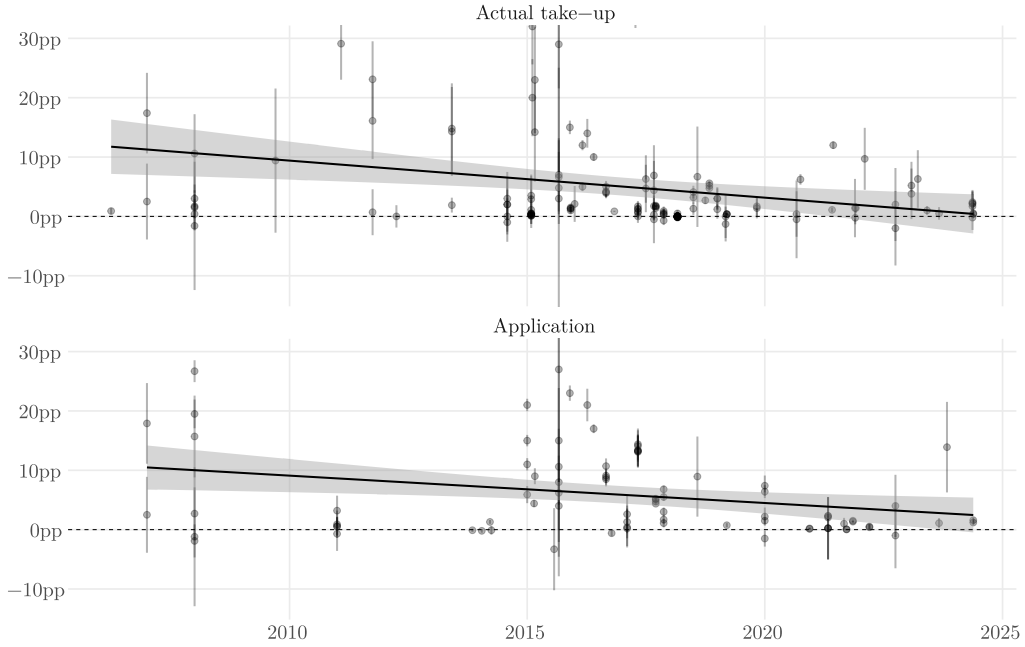


Figure 28: Estimated effects by the year of the study. Each point is one effect size with its confidence interval; the line is a linear trend with its 95% confidence band. Upper panel: studies measuring actual take-up. Lower panel: studies measuring application.

9.5 Discussion and conclusion

This meta-analysis brings together 233 treatment effects from 66 field experiments that set out to raise participation in social benefit programs and estimates an overall significant positive effect. For the interventions reducing learning demands, which are generally comparable to nudges, the estimated average effect on increasing actual take-up was 3.17 percentage points in the model including all interventions and 1.65 percentage points in the US-only model. This is comparable to the estimates of information nudges' effectiveness from Milkman et al. (2021), where the authors estimate an average effect of 2.1 percentage points on vaccine uptake. The estimates are, however, slightly above those from DellaVigna and Linos (2022), who find that nudges generally have an effect of 1.5 percentage points when correcting for publication bias. This difference could be explained by publication bias or the difference between studying social benefits and other types of behavior, such as increasing public survey response rates or police recruit test attendance. For the studies not comparable to nudges, the above models suggest that the predicted average effect size for reducing compliance demands is 15.13 percentage points for studies measuring application and 7.32 percentage points for studies measuring actual take-up.

While there are limitations to meta-analyses highlighted throughout, the estimates provide a summary of a broad literature and a comparison of shared characteristics. Across a wide range of benefits, countries, and intervention designs, three broad lessons can be drawn. First, across the globe there are problems of low take-up of social benefits, as the analyses suggest that it is possible to raise take-up of many different programs in many different countries. Second, interventions that stop at the application stage might paint an overly optimistic picture.

It is not clear whether the difference between the estimated effect sizes for the two types of outcomes is due to people failing eligibility checks or subsequent steps in the application process; nevertheless, this suggests that large application effects should be treated with caution. More research incorporating the differences in rejection rates across treatment groups could provide clearer answers.

Third, compared to easing learning demands, reducing compliance demands delivers the largest effect on take-up. This suggests that in order to combat low take-up, processes should either be simplified or access to assistance should be made easier, as it might not be enough to simply inform people of their eligibility. However, these interventions are more expensive and correcting for the cost of interventions might actually suggest that simple outreach campaigns are more efficient as all models showed that these have a positive effect, averaging across the different contexts. Similarly, there might also be a difference in the actual CATEs for the studies, as it is more likely that people are receiving the treatment in the compliance demand reducing studies, whereas it might not be everyone who has actually received the various outreach materials in the studies reducing learning demands. Although this presumably will not account for the entire difference between the two estimates, for future studies it is worth considering how much of the ITT estimates are reduced by treatment non-compliers.

Using the administrative-burden framework helps organize the disparate literature on social benefit take-up. Recoding every treatment as either a reduction in learning demands or a reduction in compliance demands makes the results from studies of housing vouchers, tax credits, food assistance, and student aid more comparable to each other and to the rest of the administrative burden literature. Collectively, the findings align with the broader idea within this literature that the more work residents must do themselves, the harder it is for them to interact with government and receive the benefits they are entitled to (Herd et al., 2013; Herd and Moynihan, 2018). These demands can arise from many verification steps or from having to learn and comply with specific program rules and reducing them would help to raise participation in government programs and help do what they are meant to do: support vulnerable people.

Future work might incorporate more studies. While 66 studies comprise a substantial body of evidence, the number is still somewhat limited, and not all of the heterogeneity across countries and benefits can be modeled precisely. Including studies on the take-up of job training, cancer screenings, vaccinations, or other services that are not directly related to income, or including quasi-experimental studies, which generally report larger effects, would slightly help this problem; however, this would affect the measurement harmonization of the meta-analysis. Furthermore, most of the studies are from the United States. While the mechanisms involved in applying for a benefit include some form of learning and compliance demand, welfare politics differs around the world. More well-powered field experiments in contexts outside the United States would help to test whether the patterns found here replicate, uncover new barriers, and show which design choices matter most across welfare regimes.

9.6 Data availability

The data can be accessed at Harvard Dataverse.

9.7 Acknowledgements

I thank the editors of JPAM, three anonymous reviewers, Jessica Lasky-Fink, Mogens Jin Pedersen, Asmus Leth Olsen, Jonas Krogh Madsen, Simon Calmar Andersen, Martin Bækgaard and Jesper Asring Hansen for their helpful comments and suggestions.

References

Adam, Christian (2025). "More Policies, More Work? An Epidemiological Assessment of Accumulating Implementation Stress in the Context of German Pension Policy". In: *Regulation & Governance* 19.3, pp. 607–617.

Adam, Christian et al. (2019). *Policy Accumulation and the Democratic Responsiveness Trap*. Cambridge: Cambridge University Press.

Adam, Christian et al. (2021). "Differential Discrimination Against Mobile EU Citizens: Experimental Evidence from Bureaucratic Choice Settings". In: *Journal of European Public Policy* 28.5, pp. 742–760.

Adams, Ian T., Scott M. Mourtgos, and Sharon H. Mastracci (2021). "High-Stakes Administrative Discretion: What Drives Body-Worn Camera Activations?" In: *Public Administration Review* 81.4, pp. 691–703.

Adams-Prassl, Abi et al. (2024). "Violence Against Women at Work". In: *The Quarterly Journal of Economics* 139.2, pp. 937–991.

Agan, Amanda, Jennifer Doleac, and Anna Harvey (2023). "Misdemeanor Prosecution". In: *The Quarterly Journal of Economics* 138.3, pp. 1453–1505.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Cambridge, MA: Harvard Business Press.

Aizer, Anna (Aug. 2007). "Public Health Insurance, Program Take-Up, and Child Health". In: *The Review of Economics and Statistics* 89.3, pp. 400–415.

Aizer, Anna and Joseph J. Doyle (2015). "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges". In: *Quarterly Journal of Economics* 130.2, pp. 759–803.

Aizer, Anna et al. (Apr. 2016). "The Long-Run Impact of Cash Transfers to Poor Families". In: *American Economic Review* 106.4, pp. 935–971.

Akata, Zeynep et al. (2020). "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence". In: *Computer* 53.8, pp. 18–28.

Alatas, Vivi et al. (Apr. 2016). "Self-Targeting: Evidence from a Field Experiment in Indonesia". In: *Journal of Political Economy* 124.2, pp. 371–427.

Alcadipani, Rafael, Gabriela Lotta, and Nissim Cohen (2024). "Natural born violence? Understanding street-level bureaucrats' use of violence: Police officers and protests". In: *Public Administration Review* 84.1, pp. 100–114.

Aleksovska, Marija, Thomas Schillemans, and Stephan Grimmelikhuijsen (2022). "Management of Multiple Accountabilities Through Setting Priorities: Evidence from a Cross-National Conjoint Experiment". In: *Public Administration Review* 82.1, pp. 132–146.

Alexeev, Sergey and Don Weatherburn (2022). "Fines for illicit drug use do not prevent future crime: evidence from randomly assigned judges". In: *Journal of Economic Behavior & Organization* 200, pp. 555–575.

Alon-Barkat, Saar and Madalina Busuioc (2023). "Human–AI Interactions in Public Sector Decision Making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice". In: *Journal of Public Administration Research and Theory* 33.1, pp. 153–169.

Alon-Barkat, Saar et al. (2025). "Algorithmic Discrimination in Public Service Provision: Understanding Citizens' Attribution of Responsibility for Human versus Algorithmic Discriminatory Outcomes". In: *Journal of Public Administration Research and Theory* 35.4, pp. 469–88.

Amaral, Sofia et al. (2023). *Deterrence or Backlash? Arrests and the Dynamics of Domestic Violence*. Working Paper 30855. National Bureau of Economic Research.

Amirapu, Amrit et al. (2022). *Personalized information provision and the take-up of emergency government benefits: Experimental evidence from India*. Tech. rep. School of Economics Discussion Papers.

Amirkhanyan, Anna A. (2010). "Monitoring across Sectors: Examining the Effect of Nonprofit and For-Profit Contractor Ownership on Performance Monitoring in State and Local Contracts". In: *Public Administration Review* 70.5, pp. 742–755.

Andersen, Simon Calmar and Thorbjørn Sejr Guul (2019). "Reducing Minority Discrimination at the Front Line—Combined Survey and Field Experimental Evidence". In: *Journal of Public Administration Research and Theory* 29.3, pp. 429–444.

Anderson, James M., Jeffrey R. Kling, and Kate Stith (1999). "Measuring Interjudge Sentencing Disparity: Before and After the Federal Sentencing Guidelines". In: *Journal of Law and Economics* 42.S1, pp. 271–308.

Anderson, Jonathan (1983). "Lix and Rix: Variations on a little-known readability index". In: *Journal of Reading* 26.6, pp. 490–496.

Andrews, Isaiah and Maximilian Kasy (2019). "Identification of and Correction for Publication Bias". In: *American Economic Review* 109.8, pp. 2766–2794.

Aneja, Abhay P. and Carlos F. Avenancio-León (2020). "No Credit for Time Served? Incarceration and Credit-Driven Crime Cycles". Working paper.

Angelova, Victoria, Will Dobbie, and Crystal S. Yang (2026). "Algorithmic Recommendations and Human Discretion". In: *Review of Economic Studies* 93.4, pp. 2250–2283.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables". In: *Journal of the American Statistical Association* 91.434, pp. 444–455.
Antill, Samuel (2022). "Do the right firms survive bankruptcy?" In: *Journal of Financial Economics* 144.2, pp. 523–546.
Araujo, Aloisio et al. (2023). "The Labor Effects of Judicial Bias in Bankruptcy". In: *Journal of Financial Economics* 150.2, p. 103720.
Arbour, William and Steeve Marchand (2022). *Parole, Recidivism, and the Role of Supervised Transition*. Working Paper 725. University of Toronto, Department of Economics.
Arkes, Hal R. (1991). "Costs and Benefits of Judgment Errors: Implications for Debiasing". In: *Psychological Bulletin* 110.3, pp. 486–98.
Armour, Philip (Aug. 2018). "The Role of Information in Disability Insurance Application: An Analysis of the Social Security Statement Phase-In". In: *American Economic Journal: Economic Policy* 10.3, pp. 1–41.
Arnold, David, Will Dobbie, and Crystal S. Yang (2018). "Racial Bias in Bail Decisions". In: *The Quarterly Journal of Economics* 133.4, pp. 1885–1932.
Arteaga, Carolina (2023). "Parental Incarceration and Children's Educational Attainment". In: *Review of Economics and Statistics* 105.6, pp. 1394–1410.
Ash, Elliott, Christoph Goessmann, and Suresh Naidu (2024). "Scaling laws: legal and social complexity in US localities". In: *Philosophical Transactions of the Royal Society A* 382.2270, p. 20230151.
Ash, Elliott, Massimo Morelli, and Matia Vannoni (2025). "More laws, more growth? Evidence from US states". In: *Journal of Political Economy* 133.7, pp. 2139–2179.
Asuming, Patrick Opoku, Hyuncheol Bryant Kim, and Armand Sim (2024). "Selection and Behavioral Responses of Health Insurance Subsidies in the Long Run: Evidence from a Field Experiment in Ghana". In: *Health Economics* 33.5, pp. 992–1032.
Athey, Susan (2017). "Beyond Prediction: Using Big Data for Policy Problems". In: *Science* 355.6324, pp. 483–85.
Augustine, Elsa et al. (2022). "The Impact of Felony Diversion in San Francisco". In: *Journal of Policy Analysis and Management* 41.3, pp. 683–709.
Autor, David et al. (2019). "Disability Benefits, Consumption Insurance, and Household Labor Supply". In: *American Economic Review* 109.7, pp. 2613–2654.
Avery, Christopher et al. (2021). "Digital messaging to improve college enrollment and success". In: *Economics of Education Review* 84, p. 102170.
Axbard, Sebastian (2024). "Convicting Corrupt Officials: Evidence from Randomly Assigned Cases". In: *Review of Economic Studies* 92.4, pp. 2271–2302.
Axt, Jordan R. and Calvin K. Lai (2019). "Reducing Discrimination: A Bias Versus Noise Perspective". In: *Journal of Personality and Social Psychology: Attitudes and Social Cognition* 117.1, pp. 26–49.
Backhaus, Leonie and Rick Vogel (2022). "Leadership in the public sector: A meta-analysis of styles, outcomes, contexts, and methods". In: *Public Administration Review* 82.6, pp. 986–1003.
Bækgaard, Martin and Tara Tankink (2022). "Administrative burden: Untangling a bowl of conceptual spaghetti". In: *Perspectives on Public Management and Governance* 5.1, pp. 16–21.
Bai, Chong-En et al. (2021). "Boosting Pension Enrollment and Household Consumption by Example: A Field Experiment on Information Provision". In: *Journal of Development Economics* 150, p. 102622.
Baillon, Aurélien et al. (2022). "Persistent effects of temporary incentives: Evidence from a nationwide health insurance experiment". In: *Journal of Health Economics* 81, p. 102580.
Baker, Daniel Brice and Shahidul Hassan (2021). "Gender and Prosecutorial Discretion: An Empirical Assessment". In: *Journal of Public Administration Research and Theory* 31.1, pp. 73–90.
Bakx, Pieter et al. (2020). "Better off at home? Effects of nursing home eligibility on costs, hospitalizations and survival". In: *Journal of Health Economics* 73, p. 102354.
Balasubramanian, Natarajan, Yang Ye, and Mingtao Xu (2022). "Substituting Human Decision-Making with Machine Learning: Implications for Organizational Learning". In: *Academy of Management Review* 47.3, pp. 448–65.
Bald, Anthony et al. (2022). "The Causal Impact of Removing Children from Abusive and Neglectful Homes". In: *Journal of Political Economy* 130.7, pp. 1919–1962.
Ban, Pamela and Seth J. Hill (2025). "Efficacy of Congressional Oversight". In: *American Political Science Review* 119.4, pp. 1742–1760.
Banerjee, Abhijit et al. (2021). "The challenges of universal health insurance in developing countries: experimental evidence from Indonesia's National Health Insurance". In: *American Economic Review* 111.9, pp. 3035–3063.
Bargain, Olivier, Herwig Immervoll, and Heikki Viitamäki (2012). "No claim, no pain. Measuring the non-take-up of social assistance using register data". In: *The Journal of Economic Inequality* 10, pp. 375–395.
Barocas, Solon and Andrew D. Selbst (2016). "Big Data's Disparate Impact". In: *California Law Review* 104.3, pp. 671–732.
Baron, E Jason and Max Gross (2025). "Is there a foster care-to-prison pipeline? Evidence from quasi-randomly assigned investigators". In: *Review of Economics and Statistics*, pp. 1–46.

Baviskar, Siddhartha and Søren C. Winter (2017). "Street-Level Bureaucrats as Individual Policymakers: The Relationship between Attitudes and Coping Behavior toward Vulnerable Children and Youth". In: *International Public Management Journal* 20.2, pp. 316–353.

Beam, Craig A., Peter M. Layde, and Daniel C. Sullivan (1996). "Variability in the Interpretation of Screening Mammograms by US Radiologists: Findings from a National Sample". In: *Archives of Internal Medicine* 156.2, pp. 209–213.

Beasley, Ryan A. and Joanna N. Lahey (2024). "Correspondence Audit Studies in Public Administration: A Systematic Review with Commentary". In: *Journal of Behavioral Public Administration* 7.1, pp. 1–30.

Beck Jørgensen, Torben (2007). "Public Values, Their Nature, Stability and Change. The Case of Denmark". In: *Public Administration Quarterly* 30.4, pp. 365–398.

Beck Jørgensen, Torben and Barry Bozeman (2007). "Public Values: An Inventory". In: *Administration & Society* 39.3, pp. 354–381.

Beecroft, Erik (May 2012). *EITC Take-Up by Recipients of Public Assistance in Virginia, and Results of a Low-Cost Experiment to Increase EITC Claims*. Report. Virginia Department of Social Services.

Beetham, David (2013). *The Legitimation of Power*. 2nd. Basingstoke: Palgrave Macmillan.

Behn, Robert D. (2001). *Rethinking Democratic Accountability*. Washington, DC: Brookings Institution Press.

Bell, Elizabeth and Sebastian Jilke (2024). "Racial Discrimination and Administrative Burden in Access to Public Services". In: *Scientific Reports* 14, p. 1071.

Bell, Elizabeth and Katharine Meyer (2024). "Does Reducing Street-Level Bureaucrats' Workload Enhance Equity in Program Access? Evidence from Burdensome College Financial Aid Programs". In: *Journal of Public Administration Research and Theory* 34.1, pp. 16–38.

Bendtsen, Karl-Emil (2026a). "Increasing Take-Up of Social Benefits: A Meta-Analysis of Field Experiments". In: *Journal of Policy Analysis and Management* 45.2, e70085.

— (2026b). *Udviklingen i den formelle regulering af de danske kommuner, 1986–2025*. Bilagsnotat. Magtudredningen 2.0.

Bendtsen, Karl-Emil and Jonas Krogh Madsen (2025). "Longer Laws, Better Service? Legislative Complexity and Bureaucratic Quality". Working paper.

Bendtsen, Karl-Emil, Lene Holm Pedersen, and Frederik Klaaborg Kjøller (2026). "Womens causal impact through the policy process: Different preferences–same outcomes". In: *Electoral Studies* 102, p. 103106.

Berggren, Andrea (2022). *Antibiotic Consumption and Health Care Utilization in Children*. Working Paper 4241676. SSRN.

Bergman, Peter, Jeffrey T Denning, and Dayanand Manoli (2019). "Is information enough? The effect of information about education tax benefits on student outcomes". In: *Journal of Policy Analysis and Management* 38.3, pp. 706–731.

Bernstein, Shai, Emanuele Colonnelli, and Benjamin Iverson (2019). "Asset Allocation in Bankruptcy". In: *The Journal of Finance* 74.1, pp. 5–53.

Bertelli, Anthony M. and Norma M. Riccucci (2022). "What Is Behavioral Public Administration Good For?" In: *Public Administration Review* 82.1, pp. 179–83.

Bertrand, Marianne and Sendhil Mullainathan (2004). "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination". In: *American Economic Review* 94.4, pp. 991–1013.

Best, Michael Carlos, Jonas Hjort, and David Szakonyi (2023). "Individuals and Organizations as Sources of State Effectiveness". In: *American Economic Review* 113.8, pp. 2121–2167.

BETA, Behavioural Economics Team of the Australian Government (Dec. 2017). *Effective Use of SMS: Timely Reminders to Report on Time*.

Bettinger, Eric P et al. (2012). "The role of application assistance and information in college decisions: Results from the H&R Block FAFSA experiment". In: *The Quarterly Journal of Economics* 127.3, pp. 1205–1242.

Bhargava, Saurabh and Dayanand Manoli (Nov. 2015). "Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment". In: *American Economic Review* 105.11, pp. 3489–3529.

Bhuller, Manudeep and Henrik Sigstad (2022). "Errors and Monotonicity in Judicial Decision-Making". In: *Economics Letters* 215, p. 110486.

Bhuller, Manudeep et al. (2018). *Incarceration Spillovers in Criminal and Family Networks*. Working Paper 24878. National Bureau of Economic Research.

Bhuller, Manudeep et al. (2020). "Incarceration, Recidivism, and Employment". In: *Journal of Political Economy* 128.4, pp. 1269–1324.

Binns, Reuben (2022). "Human Judgment in Algorithmic Loops: Individual Justice and Automated Decision-Making". In: *Regulation & Governance* 16.1, pp. 197–211.

Bird, Kelli A et al. (2021). "Nudging at scale: Experimental evidence from FAFSA completion campaigns". In: *Journal of Economic Behavior & Organization* 183, pp. 105–128.

Björnsson, Carl-Hugo (1968). *Läsbarhet: hur skall man som författare nå fram till läsarna?* Bokförlaget Liber.

Black, Bernard et al. (2024). "The effect of disability insurance receipt on mortality". In: *Journal of Public Economics* 229, p. 105033.

Blæhr, Emely Ek and Rikke Søgaard (2021). "Instrumental variable-based assessment of the effect of psychotherapy on suicide attempts, health, and economic outcomes in schizophrenia". In: *Health Economics* 30.4, pp. 903–914.
Boer, Noortje de, Jasper Eshuis, and Erik-Hans Klijn (2018). "Does Disclosure of Performance Information Influence Street-level Bureaucrats' Enforcement Style?" In: *Public Administration Review* 78.5, pp. 694–704.
Boittin, Margaret L., Rachel S. Fisher, and Cecilia Hyunjung Mo (2024). "Evidence of Caste-Class Discrimination from a Conjoint Analysis of Law Enforcement Officers". In: *American Political Science Review* 118.1, pp. 504–511.
Bolhaar, Jonneke, Nadine Ketel, and Bas van der Klaauw (2020). "Caseworker's Discretion and the Effectiveness of Welfare-to-Work Programs". In: *Journal of Public Economics* 183, p. 104080.
Bonaccio, Silvia and Reeshad S. Dalal (2006). "Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences". In: *Organizational Behavior and Human Decision Processes* 101.2, pp. 127–151.
Bonezzi, Andrea, Massimiliano Ostinelli, and Johann Melzner (2022). "The Human Black-Box: The Illusion of Understanding Human Better than Algorithmic Decision-Making". In: *Journal of Experimental Psychology: General* 151.9, pp. 2250–58.
Bonfim, Diana and Gil Nogueira (2024). *Corporate Reorganization and the Reallocation of Labor in Bankruptcy*. Working Paper 2024/2974. European Central Bank.
Bos, Marieke, Andrew Hertzberg, and Andres Liberman (2023). *The Effects of Diagnosing a Young Adult with a Mental Illness: Evidence from Randomly Assigned Doctors*. Research Paper 23-11. Swedish House of Finance.
Bovens, Mark and Stavros Zouridis (2002). "From Street-Level to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control". In: *Public Administration Review* 62.2, pp. 174–184.
Bozeman, Barry (1993). "A theory of government 'red tape'". In: *Journal of Public Administration Research and Theory* 3.3, pp. 273–304.
— (2007). *Public Values and Public Interest: Counterbalancing Economic Individualism*. Washington, DC: Georgetown University Press.
Braithwaite, John and Valerie Braithwaite (1995). "The Politics of Legalism: Rules versus Standards in Nursing-Home Regulation". In: *Social & Legal Studies* 4.3, pp. 307–341.
Brehm, John and Scott Gates (1997). *Working, Shirking, and Sabotage: Bureaucratic Response to a Democratic Public*. Ann Arbor: University of Michigan Press.
Brighton, Henry and Gerd Gigerenzer (2015). "The Bias Bias". In: *Journal of Business Research* 68.8, pp. 1772–1784.
Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2020). "Methods matter: P-hacking and publication bias in causal analysis in economics". In: *American Economic Review* 110.11, pp. 3634–3660.
Brodkin, Evelyn Z. (1997). "Inside the Welfare Contract: Discretion and Accountability in State Welfare Administration". In: *Social Service Review* 71.1, pp. 1–33.
— (2000). *Investigating Policy's 'Practical' Meaning: Street-Level Research on Welfare Policy*. JCPR Working Paper 162. Chicago: Northwestern University/University of Chicago Joint Center for Poverty Research.
— (2011). "Policy Work: Street-Level Organizations Under New Managerialism". In: *Journal of Public Administration Research and Theory* 21.suppl_2, pp. i253–i277.
Brunswik, Egon (1955). "Representative Design and Probabilistic Theory in a Functional Psychology". In: *Psychological Review* 62.3, pp. 193–217.
Bullock, Justin B. (2019). "Artificial Intelligence, Discretion, and Bureaucracy". In: *American Review of Public Administration* 49.7, pp. 751–61.
Bullock, Justin B., Matthew M. Young, and Yi-Fan Wang (2020). "Artificial Intelligence, Bureaucratic Form, and Discretion in Public Service". In: *Information Polity* 25.4, pp. 491–506.
Burton, Jason W., Mari-Klara Stein, and Tina B. Jensen (2020). "A Systematic Review of Algorithm Aversion in Augmented Decision Making". In: *Journal of Behavioral Decision Making* 33.2, pp. 220–39.
Busch, Peter A. and Helle Z. Henriksen (2018). "Digital Discretion: A Systematic Literature Review of ICT and Street-Level Discretion". In: *Information Polity* 23.1, pp. 3–28.
Bushway, Shawn D., Emily G. Owens, and Anne Morrison Piehl (2012). "Sentencing Guidelines and Judicial Discretion: Quasi-Experimental Evidence from Human Calculation Errors". In: *Journal of Empirical Legal Studies* 9.2, pp. 291–319.
Busuioc, Madalina (2021). "Accountable Artificial Intelligence: Holding Algorithms to Account". In: *Public Administration Review* 81.5, pp. 825–836.
Byrne, William Hamilton et al. (2024). "Data-Driven Futures of International Refugee Law". In: *Journal of Refugee Studies* 37.4, pp. 915–937.
Cahuc, Pierre et al. (2024). "Judge Bias in Labor Courts and Firm Performance". In: *Journal of the European Economic Association* 22.3, pp. 1319–1366.
Calvert, Randall L., Mathew D. McCubbins, and Barry R. Weingast (1989). "A Theory of Political Control and Agency Discretion". In: *American Journal of Political Science* 33.3, pp. 588–611.
Cantarelli, Paolo, Nicola Belle, and Jeremy L. Hall (2023). "Information Use in Public Administration and Policy Decision-Making: A Research Synthesis". In: *Public Administration Review* 83.6, pp. 1667–86.

Carbuccia, Laudine et al. (2025). "A Randomized Controlled Trial on the Effect of Administrative Burden and Information Costs on Social Inequalities in Early Childcare Access in France". In: *Nature Human Behaviour* 10.1, pp. 64–79.
Cárdenas, Sergio and Edgar E. Ramírez de la Cruz (2017). "Controlling Administrative Discretion Promotes Social Equity? Evidence from a Natural Experiment". In: *Public Administration Review* 77.1, pp. 80–89.
Carlsson, Hanna, Ludo Glimmerveen, and E. Lianne Visser (2026). "Between breaking and restoring boundaries: conceptualizing responsiveness in street-level decision-making". In: *Journal of Public Administration Research and Theory* 36.1, pp. 129–143.
Carpenter, Bob et al. (2017). "Stan: A Probabilistic Programming Language". In: *Journal of Statistical Software* 76.1, pp. 1–32.
Caskurlu, Tolga (2019). *Effects of Patent Rights on Acquisitions and Small Firm R&D*. Working Paper 2427191. SSRN.
Castell, Laura et al. (2022). *Take-up of Social Benefits: Experimental Evidence from France*. Tech. rep. University of Warwick.
Cecchini, Mathilde and Gitte Sommer Harrits (2022). "The Professional Agency Narrative—Conceptualizing the Role of Professional Knowledge in Frontline Work". In: *Journal of Public Administration Research and Theory* 32.1, pp. 41–57.
Cespedes, Jacelly, Carlos Parra, and Clemens Sialm (2025). "The Effect of Principal Reduction on Household Distress: Evidence from Mortgage Cramdown". In: *The Review of Financial Studies* 39.2, pp. 518–561.
Chan, David, David Card, and Lowell Taylor (2023). "Is There a VA Advantage? Evidence from Dually Eligible Veterans". In: *American Economic Review* 113.11, pp. 3003–3043.
Chan, David C., Matthew Gentzkow, and Chuan Yu (2022). "Selection with Variation in Diagnostic Skill: Evidence from Radiologists". In: *Quarterly Journal of Economics* 137.2, pp. 729–783.
Chang, Ahrum (2026). "Recurring Street-Level Encounters: How Bureaucratic Representation Changes Through Citizen Interactions". In: *Public Administration Review* 86.2, pp. 341–353.
Chanley, Virginia A., Thomas J. Rudolph, and Wendy M. Rahn (2000). "The Origins and Consequences of Public Trust in Government: A Time Series Analysis". In: *Public Opinion Quarterly* 64.3, pp. 239–56.
Charrow, Robert P. and Veda R. Charrow (1979). "Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions". In: *Columbia Law Review* 79.7, pp. 1306–1374.
Chatterjee, Sheshadri, Sangeeta Khorana, and Hatice Kizgin (2022). "Harnessing the Potential of Artificial Intelligence to Foster Citizens' Satisfaction: An Empirical Study on India". In: *Government Information Quarterly* 39.4, Article 101621.
Chen, Daniel L. and Markus Loecher (2025). "Mood and the Malleability of Moral Reasoning: The Impact of Irrelevant Factors on Judicial Decisions". In: *Journal of Behavioral and Experimental Economics* 116, p. 102364.
Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014). "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates". In: *American Economic Review* 104.9, pp. 2593–2632.
Choi, Donghyun Danny, J. Andrew Harris, and Fiona Shen-Bayh (2022). "Ethnic Bias in Judicial Decision Making: Evidence from Criminal Appeals in Kenya". In: *American Political Science Review* 116.3, pp. 1067–1080.
Choudhary, Vivek et al. (2025). "Human-AI ensembles: When can they work?" In: *Journal of Management* 51.2, pp. 536–569.
Chyn, Eric, Brigham Frandsen, and Emily Leslie (2025). "Examiner and Judge Designs in Economics: A Practitioner's Guide". In: *Journal of Economic Literature* 63.2, pp. 401–439.
Cid, Alejandro et al. (2019). "Strategies to increase the take-up of social benefits. Evidence from a field experiment in a deeply vulnerable population". In: *Documentos de trabajo del Departamento de Economía*.
Clancy, Kevin et al. (1981). "Sentence Decisionmaking: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity". In: *Journal of Criminal Law and Criminology* 72.2, pp. 524–554.
Coglianese, Cary and David Lehr (2017). "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era". In: *Georgetown Law Journal* 105, pp. 1147–1223.
Cohen, Elior (2024). "Housing the Homeless: The Effect of Placing Single Adults Experiencing Homelessness in Housing Programs on Future Homelessness and Socioeconomic Outcomes". In: *American Economic Journal: Applied Economics* 16.2, pp. 130–175.
Cohen, Nissim et al. (2023). "Varieties of welfare markets from a street-level perspective: Comparing long-term care services in Germany and Israel". In: *Public Administration Review* 83.2, pp. 419–428.
Collinson, Robert et al. (2024). "Eviction and Poverty in American Cities". In: *The Quarterly Journal of Economics* 139.1, pp. 57–120.
Compton, Mallory E et al. (2023). "Administrative Errors and Race: Can Technology Mitigate Inequitable Administrative Outcomes?" In: *Journal of Public Administration Research and Theory* 33.3, pp. 512–528.
Condorcet Marie Jean Antoine Nicolas de Caritat, Marquis de (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie Royale.
Cormen, Thomas T. et al. (2022). *Introduction to Algorithms*. 4th. Massachusetts: MIT Press.
Cortés, Tomás (2025). "Juvenile Pretrial Detention and Adult Recidivism". Master's thesis. University of Chile.

Cowgill, Bo (2019). *Bias and Productivity in Humans and Machines*. Research Paper. Posted on SSRN 21 May 2020. Columbia Business School.

Cranor, Taylor, Jacob Goldin, and Sarah Kotb (2019). "Does Informing Employees About Tax Benefits Increase Take-Up? Evidence from EITC Notification Laws". In: *National Tax Journal* 72.2, pp. 397–434.

Cruces, Guillermo Antonio et al. (2020). *Low take-up and financial inclusion: Experimental evidence from Argentina's cash transfers*. PEP Working Paper 2020-22. Partnership for Economic Policy, pp. 1–27.

Currie, Janet (May 2004). *The Take Up of Social Benefits*. Working Paper.

Currie, Janet and W. Bentley MacLeod (2017). "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians". In: *Journal of Labor Economics* 35.1, pp. 1–43.

— (2020). "Understanding Doctor Decision Making: The Case of Depression Treatment". In: *Econometrica* 88.3, pp. 847–878.

Dahl, Gordon B., Andreas Ravndal Kostøl, and Magne Mogstad (2014). "Family Welfare Cultures". In: *Quarterly Journal of Economics* 129.4, pp. 1711–1752.

Dahler-Larsen, Peter (2019). *Quality: From Plato to Performance*. Cham: Palgrave Macmillan.

Daigneault, Pierre-Marc (2024). "Reconceptualizing administrative burden around onerous experiences". In: *Perspectives on Public Management and Governance* 7.4, pp. 124–136.

Daigneault, Pierre-Marc et al. (2025). "How effective are behavioral interventions to increase the take-up of social benefits? A systematic review of field experiments". In: *Journal of Policy Analysis and Management*.

Danks, David and Alex J. London (2017). "Algorithmic Bias in Autonomous Systems". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. Melbourne, Australia: IJCAI.org, pp. 4691–97.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso (2011). "Extraneous Factors in Judicial Decisions". In: *Proceedings of the National Academy of Sciences (PNAS) of the United States of America* 108.17, pp. 6889–92.

Daugherty, Paul R. and H. James Wilson (2018). *Human + Machine: Reimagining Work in the Age of AI*. Cambridge, MA: Harvard Business Press.

Davis, Kenneth Culp (1969). *Discretionary Justice: A Preliminary Inquiry*. Baton Rouge: Louisiana State University Press.

Davis, Randall S and Sanjay K Pandey (2024). "'Feeling out' the rules: A psychological process theory of red tape". In: *Public Administration Review* 84.6, pp. 1038–1051.

Dawes, Robyn M., David Faust, and Paul E. Meehl (1989). "Clinical versus Actuarial Judgment". In: *Science* 243.4899, pp. 1668–1674.

De La Rosa, Wendy et al. (2021). "Psychological ownership interventions increase interest in claiming government benefits". In: *Proceedings of the National Academy of Sciences* 118.35, e2106357118.

DellaVigna, Stefano and Elizabeth Linos (2022). "RCTs to scale: Comprehensive evidence from two nudge units". In: *Econometrica* 90.1, pp. 81–116.

Denhardt, Janet V. and Robert E. Denhardt (2015). *The New Public Service: Serving, Not Steering*. 4th. Routledge.

Deshpande, Manasi and Michael Mueller-Smith (2022). "Does welfare prevent crime? The criminal justice outcomes of youth removed from SSI". In: *The Quarterly Journal of Economics* 137.4, pp. 2263–2307.

Di Tella, Rafael and Ernesto Schargrodsky (2013). "Criminal Recidivism after Prison and Electronic Monitoring". In: *Journal of Political Economy* 121.1, pp. 28–73.

Diamond, Rebecca, Adam Guren, and Rose Tan (2020). *The Effect of Foreclosures on Homeowners, Tenants, and Landlords*. Working Paper 27358. Revised November 2025. National Bureau of Economic Research.

Diamond, Shari Seidman and Judith N. Levi (1996). "Improving Decisions on Death by Revising and Testing Jury Instructions". In: *Judicature* 79.5, pp. 224–232.

Diaz, Ana Maria and Luz Magdalena Salas (2022). "Pretrial Detention and Conviction". In: *European Journal of Law and Economics* 53.1, pp. 1–25.

Didwania, Stephanie Holmes (2020). "The Immediate Consequences of Federal Pretrial Detention". In: *American Law and Economics Review* 22.1, pp. 24–74.

Dietrich, Brian et al. (2026). "Prioritizing Exceptional Social Needs: Experimental Evidence on the Role of Discrimination and Client Deservingness in Public Employees' and Citizens' Discretionary Behavior". In: *Public Policy and Administration* 41.1, pp. 32–54.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey (2015). "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err". In: *Journal of Experimental Psychology: General* 144.1, pp. 114–26.

— (2018). "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them". In: *Management Science* 64.3, pp. 1155–70.

Dippel, Christian, Dustin Frye, and Bryan Leonard (2025). "Property Rights without Transfer Rights: A Study of Indian Land Allotment". In: *Journal of Political Economy Microeconomics*. Just accepted.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang (2018). "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges". In: *American Economic Review* 108.2, pp. 201–240.

Dobbie, Will, Paul Goldsmith-Pinkham, and Crystal S. Yang (2017). "Consumer Bankruptcy and Financial Health". In: *The Review of Economics and Statistics* 99.5, pp. 853–869.

Dobbie, Will and Jae Song (2015). "Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection". In: *American Economic Review* 105.3, pp. 1272–1311.
Dobbie, Will et al. (2021). "Measuring Bias in Consumer Lending". In: *Review of Economic Studies*.
Dominguez, Patricio and Leopoldo Lazcano (2025). "The Effects of Pretrial Detention on Sentencing Outcomes". SOFI Working Papers in Labour Economics No. 8/2025.
Domurat, Richard, Isaac Menashe, and Wesley Yin (2021). "The role of behavioral frictions in health insurance marketplace enrollment and risk: Evidence from a field experiment". In: *American Economic Review* 111.5, pp. 1549–1574.
Döring, Matthias, Nicolas Drathschmidt, and Stine Piilgaard Porner Nielsen (2025). "It takes (at least) two to tango: Investigating interactional dynamics between clients and caseworkers in public encounters". In: *Public Administration Review* 85.2, pp. 419–435.
Doyle, Joseph J (2007). "Child Protection and Child Outcomes: Measuring the Effects of Foster Care". In: *American Economic Review* 97.5, pp. 1583–1610.
Doyle Jr., Joseph J. (2008). "Child Protection and Adult Crime: Using Investigator Assignment to Estimate Causal Effects of Foster Care". In: *Journal of Political Economy* 116.4, pp. 746–770.
Draghi, Mario (2024). *The Future of European Competitiveness. Part A: A Competitiveness Strategy for Europe*. Report. Brussels: European Commission.
Dunleavy, Patrick and Helen Margetts (2023). "Data Science, Artificial Intelligence and the Third Wave of Digital Era Governance". In: *Public Policy and Administration* 40.2, pp. 185–214.
Dworkin, Ronald (1977). *Taking Rights Seriously*. Cambridge, MA: Harvard University Press.
Ehrlich, Sean D (2011). *Access points: an institutional theory of policy bias and policy complexity*. Oxford University Press.
Eichmeyer, Sarah and Jonathan Zhang (2022). "Pathways into Opioid Dependence: Evidence from Practice Variation in Emergency Departments". In: *American Economic Journal: Applied Economics* 14.4, pp. 271–300.
— (2023). "Primary Care Providers' Influence on Opioid Use and Its Adverse Consequences". In: *Journal of Public Economics* 217, p. 104784.
Einstein, Katherine Levine and David M. Glick (2017). "Does Race Affect Access to Government Services? An Experiment Exploring Street-Level Bureaucrats and Access to Public Housing". In: *American Journal of Political Science* 61.1, pp. 100–116.
Elish, Madeleine C. (2019). "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction". In: *Engaging Science, Technology, and Society* 5, pp. 40–60.
Emanuel, Natalia, Pim Welle, and Valentin Bolotnyy (2025). *A Danger to Self and Others: Health and Criminal Consequences of Involuntary Hospitalization*. Staff Report 1158. Federal Reserve Bank of New York.
Enamorado, Ted, Anne McDonough, and Tali Mendelberg (2024). "The Shadow Carceral State and Racial Inequality in Turnout". In: *British Journal of Political Science* 54.4, pp. 1318–1339.
Engster, Daniel (2020). "A Public Ethics of Care for Policy Implementation". In: *American Journal of Political Science* 64.3, pp. 621–633.
Engstrom, David F. et al. (2020). *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. Tech. rep. Washington, DC: Administrative Conference of the United States.
Engström, Per et al. (2019). "Increasing the take-up of the housing allowance among Swedish pensioners: a field experiment". In: *International Tax and Public Finance* 26, pp. 1353–1382.
Eren, Ozkan and Naci Mocan (2018). "Emotional Judges and Unlucky Juveniles". In: *American Economic Journal: Applied Economics* 10.3, pp. 171–205.
— (2021). "Juvenile Punishment, High School Graduation, and Adult Crime: Evidence from Idiosyncratic Judge Harshness". In: *Review of Economics and Statistics* 103.1, pp. 34–47.
Ericson, Keith Marzilli et al. (2023). *Reducing administrative barriers increases take-up of subsidized health insurance coverage: evidence from a field experiment*. Tech. rep. National Bureau of Economic Research.
Eubanks, Virginia (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: Picador, St. Martin's Press.
Famulari, Melissa (2012). *Identifying the Causal Effects of Remedial Writing on Collegiate Performance Using Placement Exam Grader Toughness*. Working Paper. University of California, San Diego.
Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist (2016). *The Bright Side of Patents*. Working Paper 21959. National Bureau of Economic Research.
Feher, Andrew et al. (2023). "Personalized Letters and Emails Increased Marketplace Enrollment Among Households Eligible for Zero-Premium Plans". In: *Health Affairs* 42.4, pp. 585–593.
Fernández-i-Marín, Xavier et al. (2024). "Bureaucratic Quality and the Gap Between Implementation Burden and Administrative Capacities". In: *American Political Science Review* 118.3, pp. 1240–1260.
Fernández-i-Marín, Xavier et al. (2025). "Testing Theories of Policy Growth: Public Demands, Interest Group Politics, Electoral Competition, and Institutional Fragmentation". In: *Journal of European Public Policy* 32.3, pp. 784–809.
Fernández-i-Marín, Xavier et al. (July 2024). "Bureaucratic overburdening in advanced democracies". In: *Public Administration Review* 84.4, pp. 696–709.

Ferraro, Vicente, Gabriela Lotta, and Mykhailo Honchar (2026). "How wars impact public administration and street-level bureaucracy: teachers and education professionals on the frontlines of the Russian occupation in Ukraine". In: *Journal of Public Administration Research and Theory* 36.1, pp. 1–18.
Fine Licht, Jenny de, Mattias Agerberg, and Peter Esaiasson (2022). "'It's Not Over When It's Over'—Post-Decision Arrangements and Empirical Legitimacy". In: *Journal of Public Administration Research and Theory* 32.1, pp. 183–99.
Fine Licht, Karl de and Jenny de Fine Licht (2020). "Artificial Intelligence, Transparency, and Public Decision-Making". In: *AI & Society* 35.4, pp. 917–26.
Finer, Herman (1941). "Administrative Responsibility in Democratic Government". In: *Public Administration Review* 1.4, pp. 335–350.
Finkelstein, Amy, Nathaniel Hendren, and Mark Shepard (2019). "Subsidizing health insurance for low-income adults: Evidence from Massachusetts". In: *American Economic Review* 109.4, pp. 1530–1567.
Finkelstein, Amy and Matthew J. Notowidigdo (2019). "Take-Up and Targeting: Experimental Evidence from SNAP". In: *Quarterly Journal of Economics* 134.3, pp. 1505–1556.
Fischer, Caroline and Kristina S. Weißmüller (2024). "What Determines Civil Servants' Error Response? Evidence from a Conjoint Experiment". In: *The American Review of Public Administration* 54.8, pp. 747–770.
Fischman, Joshua B. (2014). "Measuring Inconsistency, Indeterminacy, and Error in Adjudication". In: *American Law and Economics Review* 16.1, pp. 40–85.
Fischman, Joshua B. and Max M. Schanzenbach (2012). "Racial Disparities Under the Federal Sentencing Guidelines: The Role of Judicial Discretion and Mandatory Minimums". In: *Journal of Empirical Legal Studies* 9.4, pp. 729–764.
Floridi, Luciano et al. (2018). "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations". In: *Minds and Machines* 28.4, pp. 689–707.
Fogliato, Riccardo, Alexandra Chouldechova, and Zachary Lipton (2021). "The Impact of Algorithmic Risk Assessments on Human Predictions and Its Analysis via Crowdsourcing Studies". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2, pp. 1–24.
Fotuhi, Omid and Geoffrey Cohen (2016). *Increasing FAFSA Awards*. Project spotlight in *Nudging for Success: Using Behavioral Science to Improve the Postsecondary Student Journey*. San Jose Unified School District; pp. 21–22. ideas42.
Fraisse, Henri (2017). "Households Debt Restructuring: The Re-default Effects of a Debt Suspension". In: *The Journal of Law, Economics, and Organization* 33.4, pp. 686–717.
Franco, Annie, Neil Malhotra, and Gabor Simonovits (2014). "Publication bias in the social sciences: Unlocking the file drawer". In: *Science* 345.6203, pp. 1502–1505.
Frandsen, Brigham, Lars Lefgren, and Emily Leslie (2023). "Judging Judge Fixed Effects". In: *American Economic Review* 113.1, pp. 253–277.
Frankel, Marvin E. (1973). *Criminal Sentences: Law Without Order*. New York: Hill and Wang.
Frederickson, H. George (1997). *The Spirit of Public Administration*. San Francisco, CA: Jossey-Bass.
French, Eric and Jae Song (2014). "The Effect of Disability Insurance Receipt on Labor Supply". In: *American Economic Journal: Economic Policy* 6.2, pp. 291–337.
Friedman, Jerome H. (1997). "On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality". In: *Data Mining and Knowledge Discovery* 1.1, pp. 55–77.
Friedrich, Carl J. (1940). "Public Policy and the Nature of Administrative Responsibility". In: *Public Policy*. Ed. by Carl J. Friedrich and Edward S. Mason. Cambridge, MA: Harvard University Press, pp. 3–24.
Friedrich, Laurin and Steffen Eckhard (2026). "Inequality in Frontline Communication: Bureaucrats Talk Differently to Men and Women". In: *Journal of Public Administration Research and Theory* 36.2, pp. 158–172.
Fuenzalida, Javier et al. (2026). "Red tape and burnout risks in the public service: Evidence from a survey experiment of school principals". In: *Review of Public Personnel Administration* 46.1, pp. 53–84.
Fügener, Andreas et al. (2021). "Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI". In: *MIS Quarterly* 45.3, pp. 1527–56.
Fukuyama, Francis (2013). "What is governance?" In: *Governance* 26.3, pp. 347–368.
Fuller, Lon L. (1964). *The Morality of Law*. New Haven: Yale University Press.
— (1969). *The Morality of Law*. Revised. New Haven: Yale University Press.
Gailmard, Sean and John W. Patty (2007). "Slackers and Zealots: Civil Service, Policy Discretion, and Bureaucratic Expertise". In: *American Journal of Political Science* 51.4, pp. 873–889.
Galasso, Alberto and Mark Schankerman (2014). "Patents and Cumulative Innovation: Causal Evidence from the Courts". In: *The Quarterly Journal of Economics* 130.1, pp. 317–369.
Galos, Daniel R. and Alexander Coppock (2023). "Gender Composition Predicts Gender Bias: A Meta-Reanalysis of Hiring Discrimination Audit Experiments". In: *Science Advances* 9, eade7979.
Garrow, Eve E. and Oscar Grusky (2013). "Institutional Logic and Street-Level Discretion: The Case of HIV Test Counseling". In: *Journal of Public Administration Research and Theory* 23.1, pp. 103–131.
Gaudet, Frederick J., George S. Harris, and Charles W. St. John (1933). "Individual Differences in the Sentencing Tendencies of Judges". In: *Journal of Criminal Law and Criminology* 23.5, pp. 811–818.

Geman, Stuart, Elie Bienenstock, and René Doursat (1992). "Neural Networks and the Bias/Variance Dilemma". In: *Neural Computation* 4.1, pp. 1–58.
Giannella, Eric et al. (2023). *Administrative burden and procedural denials: Experimental evidence from SNAP*. Tech. rep. National Bureau of Economic Research Cambridge, MA.
Gigerenzer, Gerd and Henry Brighton (2009). "Homo Heuristicus: Why Biased Minds Make Better Inferences". In: *Topics in Cognitive Science* 1.1, pp. 107–143.
Gillooly, Jessica W. (2022). "'Lights and Sirens': Variation in 911 Call-Taker Risk Appraisal and Its Effects on Police Officer Perceptions at the Scene". In: *Journal of Policy Analysis and Management* 41.3, pp. 762–786.
Giommoni, Tommaso et al. (2026). *The Economic Costs of Ambiguous Laws*. Cambridge Working Papers in Economics 2623. Faculty of Economics, University of Cambridge.
Girth, A. M. (2014). "A Closer Look at Contract Accountability: Exploring the Determinants of Sanctions for Unsatisfactory Contract Performance". In: *Journal of Public Administration Research and Theory* 24.2, pp. 317–348.
Giulietti, Corrado, Mirco Tonin, and Michael Vlassopoulos (2019). "Racial Discrimination in Local Public Services: A Field Experiment in the United States". In: *Journal of the European Economic Association* 17.1, pp. 165–204.
Glickman, Moshe and Tali Sharot (2025). "How Human–AI Feedback Loops Alter Human Perceptual, Emotional and Social Judgements". In: *Nature Human Behaviour* 9, pp. 345–59.
Glikson, Ella and Anita W. Woolley (2020). "Human Trust in Artificial Intelligence: Review of Empirical Research". In: *Academy of Management Annals* 14.2, pp. 627–60.
Gofen, A. (2014). "Mind the Gap: Dimensions and Influence of Street-Level Divergence". In: *Journal of Public Administration Research and Theory* 24.2, pp. 473–493.
Goldin, Jacob, Ithai Z. Lurie, and Janet McCubbin (2021). "Health Insurance and Mortality: Experimental Evidence from Taxpayer Outreach". In: *The Quarterly Journal of Economics* 136.1, pp. 1–49.
Goldin, Jacob et al. (2022). "Tax filing and take-up: Experimental evidence on tax preparation outreach and benefit claiming". In: *Journal of Public Economics* 206, p. 104550.
Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár (2026). "Leniency Designs: An Operator's Manual". In: *Journal of Economic Perspectives* 40.3, pp. 213–240.
Gonçalves, Felipe M. and Steven Mello (2025). "Police Discretion and Public Safety". In: *Journal of Political Economy Microeconomics*. Just accepted.
Gong, Robin Kaiji et al. (2025). "Tickets to the Global Market: First US Patent Award and Chinese Firm Exports". In: *Journal of International Economics* 157, p. 104124.
González-Uribe, Juanita and Santiago Reyes (2021). "Identifying and boosting 'Gazelles': Evidence from business accelerators". In: *Journal of Financial Economics* 139.1, pp. 260–287.
Gooden, Susan T. (2014). *Race and Social Equity: A Nervous Area of Government*. New York: Routledge.
Goodin, Robert E. (1988). *Reasons for Welfare: The Political Theory of the Welfare State*. Princeton: Princeton University Press.
Goodnow, Frank J. (1900). *Politics and Administration: A Study in Government*. New York: Macmillan.
Gratton, Gabriele et al. (2021). "From Weber to Kafka: Political instability and the overproduction of laws". In: *American Economic Review* 111.9, pp. 2964–3003.
Grau, Nicolás, Gonzalo Marivil, and Jorge Rivera (2023). "The Effect of Pretrial Detention on Labor Market Outcomes". In: *Journal of Quantitative Criminology* 39.2, pp. 283–332.
Green, Donald P. and Daniel Winik (2010). "Using Random Judge Assignments to Estimate the Effects of Incarceration and Probation on Recidivism among Drug Offenders". In: *Criminology* 48.2, pp. 357–387.
Greer, Robert A. and Justin B. Bullock (2018). "Decreasing Improper Payments in a Complex Federal Program". In: *Public Administration Review* 78.1, pp. 14–23.
Grimmelikhuijsen, Stephan (2023). "Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making". In: *Public Administration Review* 83.2, pp. 241–62.
Grimmelikhuijsen, Stephan and Albert Meijer (2022). "Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response". In: *Perspectives on Public Management and Governance* 5.3, pp. 232–42.
Grimmelikhuijsen, Stephan et al. (2016). "Behavioral Public Administration: Combining Insights from Public Administration and Psychology". In: *Public Administration Review* 77.1, pp. 45–56.
Grimon, Marie-Pascale (2026). "Effects of the Child Protection System on Parents". In: *Journal of Human Resources*, 0424–13537R2.
Grimon, Marie-Pascale and Christopher Mills (2026). *Better Together? A Field Experiment on Human-Algorithm Interaction in Child Protection*. Working Paper 2502.08501. arXiv.
Grindaker, Morten, Andreas Ravndal Kostøl, and Kasper Roszbach (2021). *Executive Labor Market Frictions, Corporate Bankruptcy and CEO Careers*. Working Paper 15/2021. Norges Bank.
Grøn, Caroline Howard and Heidi Houlberg Salomonsen (2020). "Organizing Central Government: A Pragmatic Meritocracy?" In: *The Oxford Handbook of Danish Politics*. Ed. by Peter Munk Christiansen, Jørgen Elklit, and Peter Nedergaard. Oxford: Oxford University Press, pp. 124–140.
Grönqvist, Hans et al. (2024). *The Intergenerational Effects of Parental Incarceration*. Working Paper 1509. Research Institute of Industrial Economics.

Gross, Max and E. Jason Baron (2022). "Temporary Stays and Persistent Gains: The Causal Effects of Foster Care". In: *American Economic Journal: Applied Economics* 14.2, pp. 170–199.
Grove, William M. et al. (2000). "Clinical versus Mechanical Prediction: A Meta-Analysis". In: *Psychological Assessment* 12.1, pp. 19–30.
Gupta, Arpit, Christopher Hansman, and Ethan Frenchman (2016). "The Heavy Costs of High Bail: Evidence from Judge Randomization". In: *The Journal of Legal Studies* 45.2, pp. 471–505.
Gupta, Charu (2023). *One Product, Many Patents: Imperfect Intellectual Property Rights in the Pharmaceutical Industry*. Working Paper. Last revised April 26, 2024. Harvard University.
Gupta, Sarika (2017). "Perils of the Paperwork: The Impact of Information and Application Assistance on Welfare Program Take-Up in India". Job market paper.
Guul, Thorbjørn S. (2023). "Political Solutions to Discriminatory Behavior". In: *American Political Science Review* 117.2, pp. 719–33.
Guul, Thorbjørn Sejr, Mogens Jin Pedersen, and Niels Bjørn Grund Petersen (2021). "Creaming among Caseworkers: Effects of Client Competence and Client Motivation on Caseworkers' Willingness to Help". In: *Public Administration Review* 81.1, pp. 12–22.
Guyton, John et al. (May 2017). "Reminders and Recidivism: Using Administrative Data to Characterize Nonfilers and Conduct EITC Outreach". In: *American Economic Review* 107.5, pp. 471–475.
Haaland, Ingar, Christopher Roth, and Johannes Wohlfart (2023). "Designing information provision experiments". In: *Journal of economic literature* 61.1, pp. 3–40.
Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto (2014). "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments". In: *Political Analysis* 22.1, pp. 1–30.
Halling, Aske and Martin Bækgaard (2024). "Administrative Burden in Citizen–State Interactions: A Systematic Literature Review". In: *Journal of Public Administration Research and Theory* 34.2, pp. 180–195.
Halling, Aske and Niels Bjørn Grund Petersen (2024). "Frontline Employees' Responses to Citizens' Communication of Administrative Burdens". In: *Public Administration Review* 84.6, pp. 1017–1037.
Hammond, Kenneth R. (1996). *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. New York: Oxford University Press.
Hansen, Paw H. (2024). "Rule Bending on the Frontlines of Public Service Delivery: How and Why Caseworkers Favor the Strong". In: *International Public Management Journal* 27.3, pp. 453–66.
Harding, David J. et al. (2018). "Imprisonment and Labor Market Outcomes: Evidence from a Natural Experiment". In: *American Journal of Sociology* 124.1, pp. 49–110.
Harding, David J. et al. (2019). "A Natural Experiment Study of the Effects of Imprisonment on Violence in the Community". In: *Nature Human Behaviour* 3, pp. 671–677.
Harnisch, Michelle (Mar. 2019). *Non-Take-Up of Means-Tested Social Benefits in Germany*. SSRN Scholarly Paper. Rochester, NY.
Harrer, Mathias et al. (2021). *Doing Meta-Analysis With R: A Hands-On Guide*. 1st. Boca Raton, FL and London: Chapman & Hall/CRC Press.
Harrits, Gitte Sommer (2019). "Stereotypes in Context: How and When Do Street-Level Bureaucrats Use Class Stereotypes?" In: *Public Administration Review* 79.1, pp. 93–103.
Hart, H. L. A. (1994). *The Concept of Law*. 2nd ed. Oxford: Clarendon Press.
Headley, Andrea M., Daniel B. Baker, and Inkyu Kang (2025). "Body-worn cameras, police arrests, and bureaucratic discretion: A large-scale causal analysis across the United States". In: *Public Administration Review* 85.4, pp. 989–1003.
Headley, Andrea M. and James E. Wright (2020). "Is Representation Enough? Racial Disparities in Levels of Force and Arrests by Police". In: *Public Administration Review* 80.6, pp. 1051–1062.
Heinrich, Carolyn J (2018). "Presidential address: 'A thousand petty fortresses': Administrative burden in US immigration policies and its consequences". In: *Journal of Policy Analysis and Management* 37.2, pp. 211–239.
Helénsdotter, Ronja (2026). "Surviving Childhood: Effects of Removing a Child From Home". In: *The Review of Economic Studies* 93.2, pp. 1001–1037.
Helland, Eric and Minjae Yun (2023). "More Talk, Less Conflict: Evidence from Requiring Informal Discovery Conferences". In: *American Law and Economics Review* 25.1, pp. 129–189.
Hemesath, Sebastian and Markus Tepe (2025). "Motivated Reasoning about Artificial Intelligence in Public Policy: Comparative Evidence from Germany". In: *Behavioural Public Policy*. Published online 2025, pp. 1–21.
Hemker, Johannes and Anselm Rink (2017). "Multiple Dimensions of Bureaucratic Discrimination: Evidence from German Welfare Offices". In: *American Journal of Political Science* 61.4, pp. 786–803.
Hemmeter, Jeffrey et al. (2024). "Communicating Program Eligibility: A Supplemental Security Income (SSI) Field Experiment". Working paper, current version July 15, 2024.
Henderson, Alexander C. (2013). "Examining Policy Implementation in Health Care: Rule Abidance and Deviation in Emergency Medical Services". In: *Public Administration Review* 73.6, pp. 799–809.
Herbst, Daniel (2023). "The Impact of Income-Driven Repayment on Student Borrower Outcomes". In: *American Economic Journal: Applied Economics* 15.1, pp. 1–25.

Herd, Pamela and Donald Moynihan (2018). *Administrative Burden: Policymaking by Other Means*. New York: Russell Sage Foundation.
Herd, Pamela et al. (2013). "Shifting administrative burden to the state: The case of medicaid take-up". In: *Public Administration Review* 73.s1, S69–S81.
Hermes, Henning et al. (2025). "Application Barriers and the Socioeconomic Gap in Child Care Enrollment". In: *Journal of the European Economic Association* 23.3, pp. 1133–1172.
Hervelin, Jérémy (2022). "Directing young dropouts via SMS: evidence from a field experiment". In: *IZA Journal of Labor Policy* 12.1.
Heyes, Anthony and Soodeh Saberian (2019). "Temperature and Decisions: Evidence from 207,000 Court Cases". In: *American Economic Journal: Applied Economics* 11.2, pp. 238–265.
Highhouse, Scott (2008). "Stubborn Reliance on Intuition and Subjectivity in Employee Selection". In: *Industrial and Organizational Psychology* 1.3, pp. 333–42.
Highhouse, Scott and Margaret E. Brooks (2023). "Improving Workplace Judgments by Reducing Noise: Lessons Learned from a Century of Selection Research". In: *Annual Review of Organizational Psychology and Organizational Behavior* 10, pp. 519–533.
Hillebrand, Luis, Sebastian Raisch, and Jonathan Schad (2025). "Managing with Artificial Intelligence: An Integrative Framework". In: *Academy of Management Annals* 19.1, pp. 343–75.
Hinterleitner, Markus, Christoph Knill, and Yves Steinebach (2024). "The Growth of Policies, Rules, and Regulations: A Review of the Literature and Research Agenda". In: *Regulation & Governance* 18.2, pp. 637–654.
Hjalmarsson, Randi and Matthew J Lindquist (2019). "The Causal Effect of Military Conscription on Crime". In: *The Economic Journal* 129.622, pp. 2522–2562.
Hjortskov, Morten et al. (2026). "Implementation support: a field experiment on the effects of fidelity and professional responsibility approaches". In: *Journal of Public Administration Research and Theory* 36.2, pp. 173–187.
Ho, Daniel E. (2012). "Fudging the Nudge: Information Disclosure and Restaurant Grading". In: *Yale Law Journal* 122.3, pp. 574–688.
— (2017). "Does Peer Review Work? An Experiment of Experimentalism". In: *Stanford Law Review* 69.1, pp. 1–119.
Ho, Daniel E. and Sam Sherman (2017). "Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement". In: *Annual Review of Law and Social Science* 13, pp. 251–272.
Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li (2018). "Discretion in Hiring". In: *Quarterly Journal of Economics* 133.2, pp. 765–800.
Holland, Paul W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960.
Hollibaugh, Gary E., Matthew R. Miles, and Chad B. Newswander (2020). "Why Public Employees Rebel: Guerrilla Government in the Public Sector". In: *Public Administration Review* 80.1, pp. 64–74.
Homonoff, Tatiana and Jason Somerville (2021). "Program recertification costs: Evidence from SNAP". In: *American Economic Journal: Economic Policy* 13.4, pp. 271–298.
Honigsberg, Colleen and Matthew Jacob (2021). "Deleting Misconduct: The Expungement of BrokerCheck Records". In: *Journal of Financial Economics* 139.3, pp. 800–831.
Horsch, Philipp, Philip Longoni, and David Oesch (2021). "Intangible Capital and Leverage". In: *Journal of Financial and Quantitative Analysis* 56.2, pp. 475–498.
Hoynes, Hilary, Doug Miller, and David Simon (2015). "Income, the earned income tax credit, and infant health". In: *American Economic Journal: Economic Policy* 7.1, pp. 172–211.
Hoynes, Hilary, Diane Whitmore Schanzenbach, and Douglas Almond (2016). "Long-run impacts of childhood access to the safety net". In: *American Economic Review* 106.4, pp. 903–934.
Hubble, Cale and Chiara Varazzani (May 10, 2023). *Mapping the global behavioural insights community*. Blog post. OECD Observatory of Public Sector Innovation (OPSI).
Huber, John D. and Charles R. Shipan (2002). *Deliberate Discretion? The Institutional Foundations of Bureaucratic Autonomy*. Cambridge: Cambridge University Press.
Humlum, Anders, Jakob Munch, and Mette Rasmussen (2025). *What Works for the Unemployed? Evidence From Quasi-Random Caseworker Assignments*. Working Paper 33807. National Bureau of Economic Research.
Humlum, Anders and Emilie Vestergaard (2025). "The Unequal Adoption of ChatGPT Exacerbates Existing Inequalities among Workers". In: *Proceedings of the National Academy of Sciences (PNAS) of the United States of America* 122.1, e2414972121.
Humphries, John Eric et al. (2025). "Conviction, Incarceration, and Recidivism: Understanding the Revolving Door". In: *The Quarterly Journal of Economics* 140.4, pp. 2907–2962.
Huntington-Klein, Nick et al. (May 2025). *The Sources of Researcher Variation in Economics*. NBER Working Paper 33729. Cambridge, MA: National Bureau of Economic Research.
Hurka, Steffen, Maximilian Haag, and Constantin Kaplaner (Sept. 2022). "Policy complexity in the European Union, 1993-today: introducing the EUPLEX dataset". In: *Journal of European Public Policy* 29.9, pp. 1512–1527.

Hyman, Benjamin G. (2022). *Can Displaced Labor Be Retrained? Evidence from Quasi-Random Assignment to Trade Adjustment Assistance*. Working Paper CES-22-05. U.S. Census Bureau, Center for Economic Studies.
ideas42 (Dec. 2015). *Increasing FAFSA Applications: Making College More Affordable*. Collaboration with Arizona State University. Behavioral email intervention RCT increasing FAFSA submissions among continuing students.
ideas42 and NYC Behavioral Design Team (Jan. 2022). *A Simple Tool to Serve SNAP Clients: Increasing SNAP Recertification Rates Using Behaviorally Informed Reminders*. RCT testing behaviorally-informed reminder letters to increase SNAP recertification rates in New York City.
Imbens, Guido W. and Joshua D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". In: *Econometrica* 62.2, pp. 467–475.
Ingrams, Alex, Wesley Kaufmann, and Daan Jacobs (2022). "In AI We Trust? Citizen Perceptions of AI in Government Decision Making". In: *Policy & Internet* 14.2, pp. 390–409.
Ivanov, Ivan T. and James Z. Wang (2024). "Bank Supervision and Corporate Credit Supply". In: *Management Science* 70.5, pp. 3338–3361.
Jakobsen, Mads Leth Felsager and Peter Bjerre Mortensen (2014). *Regelstaten: Væksten i danske love og bekendtgørelser 1989–2011*. København: Jurist- og Økonomforbundets Forlag.
— (2015). "How Politics Shapes the Growth of Rules". In: *Governance* 28.4, pp. 497–515.
Jakobsen, Mette, Ann-Kristina Løkke, and Florian Keppeler (2023). "Facing the Human Capital Crisis: A Systematic Review and Research Agenda on Recruitment and Selection in the Public Sector". In: *Public Administration Review* 83.6, pp. 1635–48.
Jankowski, Michael, Christine Prokop, and Markus Tepe (2020). "Representative Bureaucracy and Public Hiring Preferences: Evidence from a Conjoint Experiment among German Municipal Civil Servants and Private Sector Employees". In: *Journal of Public Administration Research and Theory* 30.4, pp. 596–618.
Janssen, Marijn et al. (2022). "Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-Supported Decision-Making in Government". In: *Social Science Computer Review* 40.2, pp. 478–93.
Jarrahi, Mohammad H. (2018). "Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making". In: *Business Horizons* 61.4, pp. 577–86.
Jensen, Didde Cramer and Line Bjørnskov Pedersen (2017). "The Impact of Empathy—Explaining Diversity in Street-Level Decision-Making". In: *Journal of Public Administration Research and Theory* 27.3, pp. 433–449.
Jilke, Sebastian and Lars Tummers (2018). "Which Clients are Deserving of Help? A Theoretical Model and Experimental Test". In: *Journal of Public Administration Research and Theory* 28.2, pp. 226–238.
Jones, Damon (2010). "Information, preferences, and public benefit participation: Experimental evidence from the advance EITC and 401 (k) savings". In: *American Economic Journal: Applied Economics* 2.2, pp. 147–163.
Jordan, Andrew, Ezra Karger, and Derek Neal (2025). "Heterogeneous impacts of sentencing decisions". In: *Journal of Labor Economics* 43.4, pp. 1005–1034.
Jordan, Andrew and Taeho Kim (2025). "Strengthening police oversight: The impacts of misconduct investigators on police officer behavior". In: *Journal of Policy Analysis and Management* 44.4, pp. 1286–1316.
Junge, Sarah Yde (2022). "Administrative Groupings and Equality in Public Service Provision". In: *Journal of Public Administration Research and Theory* 32.2, pp. 252–268.
Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein (2021). *Noise: A Flaw in Human Judgment*. New York: Little, Brown and Company.
Kahneman, Daniel et al. (Oct. 2016). "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making". In: *Harvard Business Review*.
Kane, Thomas J. and Douglas O. Staiger (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*. Working Paper 14607. National Bureau of Economic Research.
Kang, Inkyu and Sebastian Jilke (2024). "Mapping Out the Motivational Basis of Active Representation as Intergroup Behavior". In: *Public Administration* 102.1, pp. 164–187.
Kang, Karam and Bernardo S. Silveira (2021). "Understanding Disparities in Punishment: Regulator Preferences and Expertise". In: *Journal of Political Economy* 129.10, pp. 2947–2992.
Kaplaner, Constantin and Yves Steinebach (2024). "Coping practices and the spatial dimension of authority design". In: *Public Administration Review* 84.1, pp. 86–99.
Katz, Daniel Martin and Michael J. Bommarito (2014). "Measuring the complexity of the law: the United States Code". In: *Artificial Intelligence and Law* 22.4, pp. 337–374.
Katz, Daniel Martin et al. (2020). "Complex Societies and the Growth of the Law". In: *Scientific Reports* 10, p. 18737.
Kauppi, Katri and Erik M. van Raaij (2015). "Opportunism and Honest Incompetence—Seeking Explanations for Noncompliance in Public Procurement". In: *Journal of Public Administration Research and Theory* 25.3, pp. 953–979.
Kawai, Keiichi, Ruitian Lang, and Hongyi Li (Nov. 2018). "Political Kludges". In: *American Economic Journal: Microeconomics* 10.4, pp. 131–158.
Keiser, Lael R et al. (2002). "Lipstick and Logarithms: Gender, Institutional Context, and Representative Bureaucracy". In: *American Political Science Review* 96.3, pp. 553–564.

Keiser, Lael R. (2010). "Understanding Street-Level Bureaucrats' Decision Making: Determining Eligibility in the Social Security Disability Program". In: *Public Administration Review* 70.2, pp. 247–257.
Keiser, Lael R., Peter R. Mueser, and Seung-Whan Choi (2004). "Race, Bureaucratic Discretion, and the Implementation of Welfare Reform". In: *American Journal of Political Science* 48.2, pp. 314–327.
Keiser, Lael R. and Joe Soss (1998). "With Good Cause: Bureaucratic Discretion and the Politics of Child Support Enforcement". In: *American Journal of Political Science* 42.4, pp. 1133–1156.
Kejriwal, Mayank et al. (2024). "A Noise Audit of Human-Labeled Benchmarks for Machine Commonsense Reasoning". In: *Scientific Reports* 14.1, p. 8609.
Kennedy, Ryan P., Philip D. Waggoner, and Matthew M. Ward (2022). "Trust in Public Policy Algorithms". In: *Journal of Politics* 84.2, pp. 1132–48.
Keppeler, Florian (2024). "No Thanks, Dear AI! Understanding the Effects of Disclosure and Deployment of Artificial Intelligence in Public Sector Recruitment". In: *Journal of Public Administration Research and Theory* 34.1, pp. 39–52.
Keppeler, Florian et al. (2025). "How Ensembling AI and Public Managers Improves Decision-Making". In: *Journal of Public Administration Research and Theory* 35.3, pp. 261–276.
Kilaru, Austin S. et al. (2026). "Linkage of Emergency Department Patients with Public Benefits Navigators via Text Messages". In: *JAMA Health Forum* 7.2, e256637.
Kim, Jae Yeon et al. (2025). "Administrative Checkpoints, Burdens, and Human-Centered Design: Increasing Interview Access to Raise SNAP Participation". In: *Journal of Policy Analysis and Management* 44.4, pp. 1153–1170.
Kissel, Helen and Helena Roy (2025). "Understanding Variation in Cesarean Section Use: Supply-Side Drivers and Maternal Health Effects". Unpublished manuscript.
Klaauw, Bas van der and Heike Vethaak (2022). *Empirical Evaluation of Broader Job Search Requirements for Unemployed Workers*. Discussion Paper 15698. IZA.
Kleinberg, Jon et al. (2018). "Human Decisions and Machine Predictions". In: *Quarterly Journal of Economics* 133.1, pp. 237–293.
Kleven, Henrik Jacobsen and Wojciech Kopczuk (2011). "Transfer Program Complexity and the Take-Up of Social Benefits". In: *American Economic Journal: Economic Policy* 3.1, pp. 54–90.
Knill, Christoph, Yves Steinebach, and Dionys Zink (2024). "How Policy Growth Affects Policy Implementation: Bureaucratic Overload and Policy Triage". In: *Journal of European Public Policy* 31.2, pp. 324–351.
Knuth, Donald E. (1977). "Algorithms". In: *Scientific American* 236.4, pp. 63–80.
Ko, Wonsik and Robert A. Moffitt (2024). "Take-Up of Social Benefits". In: *Handbook of Labor, Human Resources and Population Economics*. Ed. by Klaus F. Zimmermann. Cham: Springer International Publishing, pp. 1–42.
Köchling, Alina and Marius C. Wehner (2020). "Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development". In: *Business Research* 13.3, pp. 795–848.
Kogan, Vladimir (2017). "Administrative Centralization and Bureaucratic Responsiveness: Evidence from the Food Stamp Program". In: *Journal of Public Administration Research and Theory* 27.4, pp. 629–646.
Kolstad, Karoline Larsen (2026). "Overburdened Bureaucrats: Providing Equal Access to Public Services During COVID-19". In: *The Journal of Politics* 88.1, pp. 250–266.
Koppel, Stephen et al. (2024). "Examining the Causal Effect of Pretrial Detention on Case Outcomes: A Judge Fixed Effect Instrumental Variable Approach". In: *Journal of Experimental Criminology* 20.2, pp. 439–456.
Kosti, Nir (2024). "Conceptualization and measurement of regulatory discretion: Text analysis of 120 years of British legislation". In: *Regulation & Governance* 18.3, pp. 761–779.
Kuziemski, Maciej and Gianluca Misuraca (2020). "AI Governance in the Public Sector: Three Tales from the Frontiers of Automated Decision-Making in Democratic Settings". In: *Telecommunications Policy* 44.6, p. 101976.
Kyle, Michael Anne and Austin B Frakt (2021). "Patient administrative burden in the US health care system". In: *Health Services Research* 56.5, pp. 755–765.
LaForest-Tucker, Michael (2026). "Parole Supervision on the Margins". Working paper, revised February 2, 2026.
Lakens, Daniel (2017). "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses". In: *Social Psychological and Personality Science* 8.4, pp. 355–362.
Larrick, Richard P. (2004). "Debiasing". In: *Blackwell Handbook of Judgment and Decision Making*. Ed. by D. J. Koehler and N. Harvey. Malden, MA: Blackwell Publishing, pp. 316–38.
Lasky-Fink, Jessica (2023). "Essays on Participation in the Social Safety Net". PhD thesis. UC Berkeley.
Lasky-Fink, Jessica and Elizabeth Linos (2024). "Improving Delivery of the Social Safety Net: The Role of Stigma". In: *Journal of Public Administration Research and Theory* 34.2, pp. 270–283.
Lauderdale, Benjamin E, Chris Hanretty, and Nick Vivyan (2018). "Decomposing public opinion variation into ideology, idiosyncrasy, and instability". In: *The Journal of Politics* 80.2, pp. 707–712.
Lavee, Einat (2021). "Who Is in Charge? The Provision of Informal Personal Resources at the Street Level". In: *Journal of Public Administration Research and Theory* 31.1, pp. 4–20.
Lee, David S. et al. (2022). "Valid t-Ratio Inference for IV". In: *American Economic Review* 112.10, pp. 3260–3290.

Lee, Logan M. (2023). "Halfway Home? Residential Housing and Reincarceration". In: *American Economic Journal: Applied Economics* 15.3, pp. 117–149.
Lee, Yunsoo and Hindy L. Schachter (2019). "Exploring the Relationship Between Trust in Government and Citizen Participation". In: *International Journal of Public Administration* 42.5, pp. 405–16.
Lerman, Amy E, Meredith L Sadin, and Samuel Trachtman (2017). "Policy uptake as political behavior: Evidence from the affordable care act". In: *American Political Science Review* 111.4, pp. 755–770.
Leroy, Claire (Dec. 2024). *Raising Take-up of Welfare Programs: Evidence from a Large French Reform*. Working Paper. Latest version, December 2024. CREST – École polytechnique, IP Paris.
Leslie, Emily and Nolan G. Pope (2017). "The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from New York City Arraignments". In: *The Journal of Law and Economics* 60.3, pp. 529–557.
Limberg, Julian et al. (2021). "The more the better? Rule growth and policy impact from a macro perspective". In: *European Journal of Political Research* 60.2, pp. 438–454.
Linos, Elizabeth, Vikash Reddy, and Jesse Rothstein (2024). "Demystifying college costs: How nudges can and can't help". In: *Behavioural Public Policy* 8.3, pp. 497–518.
Linos, Elizabeth et al. (Nov. 2022). "Can Nudges Increase Take-Up of the EITC? Evidence from Multiple Field Experiments". In: *American Economic Journal: Economic Policy* 14.4, pp. 432–452.
Linos, Elizabeth et al. (Mar. 2025). *Interventions to Bolster Benefits Take-Up: Assessing Intensity, Framing, and Targeting of Government Outreach*. Tech. rep. RWP25-001. Working Paper. Harvard Kennedy School Faculty Research Working Paper Series.
Lipsky, Michael (1980). *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. Originally published 1980; reprinted 2010. New York: Russell Sage Foundation.
— (2010). *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. Expanded edition; originally published 1980. New York: Russell Sage Foundation.
List, John A (2022). *The voltage effect: How to make good ideas great and great ideas scale*. Crown Currency.
Liu, Hanzhang (2019). "The Logic of Authoritarian Political Selection: Evidence from a Conjoint Experiment in China". In: *Political Science Research and Methods* 7.4, pp. 853–870.
Liu, Yixin (2024). "Partisan Collaboration in Policy Adoption: An Experimental Study with Local Government Officials". In: *Policy Studies Journal* 52.4, pp. 955–967.
Løberg, Ida Bring (2023). "Assessments of Digital Client Representations: How Frontline Workers Reconstruct Client Narratives from Fragmented Information". In: *Journal of Public Administration Research and Theory* 33.1, pp. 19–29.
Loeffler, Charles E. (2013). "Does Imprisonment Alter the Life Course? Evidence on Crime and Employment from a Natural Experiment". In: *Criminology* 51.1, pp. 137–166.
Logg, Jennifer M., Julia A. Minson, and Don A. Moore (2019). "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment". In: *Organizational Behavior and Human Decision Processes* 151, pp. 90–103.
Lopes, Francisca Vargas et al. (2022). "The Effects of Supported Housing for Individuals with Mental Disorders". In: *Health Economics* 31.S2, pp. 115–133.
Lopoo, Leonard M., Colleen Heflin, and Joseph Boskovski (2020). "Testing behavioral interventions designed to improve on-time SNAP recertification". In: *Journal of Behavioral Public Administration* 3.2, pp. 1–8.
Lowande, Kenneth and Andrew Proctor (2020). "Bureaucratic Responsiveness to LGBT Americans". In: *American Journal of Political Science* 64.3, pp. 664–681.
Luan, Shenghua, Jochen Reb, and Gerd Gigerenzer (2019). "Ecological Rationality: Fast-and-Frugal Heuristics for Managerial Decision Making Under Uncertainty". In: *Academy of Management Journal* 62.6, pp. 1735–59.
Lum, Kristian, Erwin Ma, and Mike Baiocchi (2017). "The Causal Impact of Bail on Case Outcomes for Indigent Defendants in New York City". In: *Observational Studies* 3.1, pp. 38–64.
Machado Ribeiro, Beatriz (2019). "Pretrial Detention and Rearrest: Evidence from Brazil". Master's thesis. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio).
Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand (2013). "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt". In: *American Economic Review* 103.5, pp. 1797–1829.
Mahmud, Hasan et al. (2022). "What Influences Algorithmic Decision-Making? A Systematic Literature Review on Algorithm Aversion". In: *Technological Forecasting and Social Change* 175, Article 121390.
Margetts, Helen and Cosmina Dorobantu (2019). "Rethink government with AI". In: *Nature* 568.7751, pp. 163–65.
Markussen, Simen, Arnstein Mykletun, and Knut Røed (2012). "The Case for Presenteeism: Evidence from Norway's Sickness Insurance Program". In: *Journal of Public Economics* 96.11–12, pp. 959–972.
Mashaw, Jerry L. (1983). *Bureaucratic Justice: Managing Social Security Disability Claims*. New Haven: Yale University Press.
Masood, Ayesha and Muhammad Azfar Nisar (2022). "Repairing the State: Policy Repair in the Frontline Bureaucracy". In: *Public Administration Review* 82.2, pp. 256–268.
Matikka, Tuomas and Tuuli Paukkeri (2022). "Does sending letters increase the take-up of social benefits? Evidence from a new benefit program". In: *Empirical Economics* 63.6, pp. 3253–3287.
Matland, Richard E. (1995). "Synthesizing the Implementation Literature: The Ambiguity-Conflict Model of Policy Implementation". In: *Journal of Public Administration Research and Theory* 5.2, pp. 145–174.

Maynard-Moody, Steven and Michael Musheno (2003). *Cops, Teachers, Counselors: Stories from the Front Lines of Public Service*. Ann Arbor: University of Michigan Press.
— (2012). "Social Equities and Inequities in Practice: Street-Level Workers as Agents and Pragmatists". In: *Public Administration Review* 72.s1, S16–S23.
McCollough, Thomas E. (1991). *The Moral Imagination and Public Life: Raising the Ethical Question*. Chatham, NJ: Chatham House Publishers.
McDonough, Anne, Ted Enamorado, and Tali Mendelberg (2022). "Jailed While Presumed Innocent: The Demobilizing Effects of Pretrial Incarceration". In: *The Journal of Politics* 84.3, pp. 1777–1790.
McQuillan, Casey and Brendan Moore (2025). "Barriers to Benefits: Unemployment Insurance Take-Up and Labor Market Effects". Job market paper; AEA RCT Registry No. AEARCTR-0013608.
Meehl, Paul E. (1954). *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: University of Minnesota Press.
Menkveld, Albert J. et al. (2024). "Nonstandard Errors". In: *The Journal of Finance* 79.3, pp. 2339–2390.
Mikkelsen, Maria Falk, Mogens Jin Pedersen, and Niels Bjørn Grund Petersen (2023). "To Act or Not to Act? How Client Progression Affects Purposeful Performance Information Use at the Frontlines". In: *Journal of Public Administration Research and Theory* 33.2, pp. 296–312.
Milap, Akshay and Ankur Sarin (2023). "Evaluating the efficacy of demand-side communication interventions on claiming rights: evidence from an action research field experiment in India". In: *Human Communication Research* 49.1, pp. 85–103.
Milkman, Katherine L. et al. (2021). "A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment". In: *Proceedings of the National Academy of Sciences* 118.20, e2101165118.
Milkman, Kathrine L., Dolly Chugh, and Max H. Bazerman (2009). "How Can Decision Making Be Improved?" In: *Perspectives on Psychological Science* 4.4, pp. 379–83.
Miller, Gary J. and Andrew B. Whitford (2016). *Above Politics: Bureaucratic Discretion and Credible Commitment*. New York: Cambridge University Press.
Moffitt, Robert (1983). "An Economic Model of Welfare Stigma". In: *The American Economic Review* 73.5, pp. 1023–1035.
Møller, Anne Mette (2021). "Deliberation and Deliberative Organizational Routines in Frontline Decision-Making". In: *Journal of Public Administration Research and Theory* 31.3, pp. 471–488.
— (2025). "Inside the digital state: frontline work in the context of digital layering". In: *Journal of Public Administration Research and Theory* 35.2, pp. 133–147.
Møller, Anne Mette, Kirstine Zinck Pedersen, and Anja Svejgaard Pors (2022). "The Bureaucratic Ethos in Street-Level Work: Revitalizing Weber's Ethics of Office". In: *Perspectives on Public Management and Governance* 5.2, pp. 151–163.
Montjoy, Robert S. and Douglas J. Watson (1995). "A Case for Reinterpreted Dichotomy of Politics and Administration as a Professional Standard in Council-Manager Government". In: *Public Administration Review* 55.3, pp. 231–239.
Moore, Ryan T et al. (2022). "Anchor management: a field experiment to encourage families to meet critical programme deadlines". In: *Journal of Public Policy* 42.4, pp. 615–636.
Morrell, Kevin and Graeme Currie (2015). "Impossible Jobs or Impossible Tasks? Client Volatility and Frontline Policing Practice in Urban Riots". In: *Public Administration Review* 75.2, pp. 264–275.
Moseley, Alice and Eva Thomann (2021). "A Behavioural Model of Heuristics and Biases in Frontline Policy Implementation". In: *Policy & Politics* 49.1, pp. 49–67.
Mou, Hongyu, Licheng Liu, and Yiqing Xu (2023). "Panel Data Visualization in R (panelView) and Stata (panelview)". In: *Journal of Statistical Software* 107.7, pp. 1–20.
Mueller, Holger and Constantine Yannelis (2022). "Increasing enrollment in income-driven student loan repayment plans: Evidence from the Navient field experiment". In: *The Journal of Finance* 77.1, pp. 367–402.
Mullainathan, Sendhil and Ziad Obermeyer (2021). "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care". In: *The Quarterly Journal of Economics* 137.2, pp. 679–727.
Munroe, David and Laurence H. Wilse-Samson (2013). "Foreclosure Contagion: Measurement and Mechanisms". In: *Available at SSRN 4349905*.
Musgrave, Richard A. (1959). *The Theory of Public Finance: A Study in Public Economy*. New York: McGraw-Hill.
Myerson, Rebecca et al. (2022). "Personalized Telephone Outreach Increased Health Insurance Take-Up For Hard-To-Reach Populations, But Challenges Remain: Study examines personalized telephone outreach to increase take up of ACA Marketplace enrollment." In: *Health Affairs* 41.1, pp. 129–137.
Myerson, Rebecca Mary, Allison Espeseth, and Laura Dague (2025). "Navigating Medicaid: Experimental Evidence on Administrative Burden and Coverage Loss". NBER Working Paper No. 34191.
Nagin, Daniel S. and G. Matthew Snodgrass (2013). "The Effect of Incarceration on Re-Offending: Evidence from a Natural Experiment in Pennsylvania". In: *Journal of Quantitative Criminology* 29.4, pp. 601–642.
Nesbit, Rachel (2023). *The Role of Mandated Mental Health Treatment in the Criminal Justice System*. Job Market Paper. University of Maryland.
Nguyen, Ha Thi Hong et al. (2014). "Encouraging Health Insurance for the Informal Sector: A Cluster Randomized Trial". World Bank Policy Research Working Paper No. 6910.

Nichols, Albert L and Richard J Zeckhauser (1982). "Targeting transfers through restrictions on recipients". In: *The American Economic Review* 72.2, pp. 372–377.
Niskanen, William A. (1971). *Bureaucracy and Representative Government*. Chicago: Aldine-Atherton.
Norris, Samuel, Matthew Pecenco, and Jeffrey Weaver (2021). "The Effects of Parental and Sibling Incarceration: Evidence from Ohio". In: *American Economic Review* 111.9, pp. 2926–2963.
O'Neil, Cathy (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.
Oberfield, Zachary W. and Matthew B. Incantalupo (2021). "Racial Discrimination and Street-Level Managers: Performance, Publicness, and Group Bias". In: *Public Administration Review* 81.6, pp. 1055–1070.
OES, Office of Evaluation Sciences, Code for America CfA, and U.S. Department of Agriculture USDA (May 2024). *Decreasing Supplemental Nutrition Assistance Program (SNAP) Denials Due to Incomplete Information*. A randomized evaluation of a text campaign linking to a redesigned document uploader to reduce SNAP denials.
OES, Office of Evaluation Sciences and U.S. Department of Health and Human Services HHS (2024). *Increasing Access to Minnesota Child Care Stabilization Base Grants*. Texts and calls increased funding awards for child care providers in Minnesota. Evaluation period: Feb–Jun 2023. Office of Evaluation Sciences.
OES, Office of Evaluation Sciences and U.S. Department of the Treasury USDT (2022). *Increasing Take-Up of the American Opportunity Tax Credit: A Series of Emails and a Letter Increased Take-Up*. Randomized controlled trial evaluating communications to increase take-up of higher education tax benefits among undergraduate students.
— (May 2024). *Reducing Documentation Burdens to Accelerate and Broaden Access to Rental Assistance*. Evaluation of a ZIP code-based fact-specific proxy (FSP) policy to simplify income documentation for ERA in Virginia.
Ofek-Shanny, Yuval, Avner Strulov-Shlain, and Dan Zeltzer (2025). "Impacts of Home-Care Subsidies: Evidence from Quasi-Random Assignment". In: *Journal of Public Economics*.
Olsen, Asmus Leth, Karl-Emil Bendtsen, and Paul van Leeuwen (2026). "Public Administration Beyond Public Administration Journals". In: *PS: Political Science & Politics* 59.3, pp. 501–507.
Olsen, Asmus Leth, Jonas Høgh Kyhse-Andersen, and Donald Moynihan (2022). "The Unequal Distribution of Opportunity: A National Audit Study of Bureaucratic Discrimination in Primary School Access". In: *American Journal of Political Science* 66.3, pp. 587–603.
Olson, Jeremiah C. (2016). "Race and Punishment in American Prisons". In: *Journal of Public Administration Research and Theory* 26.4, pp. 758–768.
Oorschot, Wim van (1991). "Non-Take-Up of Social Security Benefits in Europe". In: *Journal of European Social Policy* 1.1, pp. 15–30.
Oprea, Ryan (2020). "What Makes a Rule Complex?" In: *American Economic Review* 110.12, pp. 3913–3951.
Osnabrügge, Moritz and Matia Vannoni (2025). "Quality of legislation and compliance: a natural language processing approach". In: *Political Science Research and Methods* 13.3, pp. 736–744.
Page, Lindsay C et al. (2023). "Financial aid nudges: A national experiment with informational interventions". In: *Educational Evaluation and Policy Analysis* 45.2, pp. 195–219.
Page, Lindsay C., Benjamin L. Castleman, and Katharine Meyer (2020). "Customized Nudging to Improve FAFSA Completion and Income Verification". In: *Educational Evaluation and Policy Analysis* 42.1, pp. 3–21.
Page, Scott E. (2007). "Making the Difference: Applying a Logic of Diversity". In: *Academy of Management Perspectives* 21.4, pp. 6–20.
Pálfi, Bence, Kavleen Arora, and Olga Kostopoulou (2022). "Algorithm-Based Advice Taking and Clinical Judgement: Impact of Advice Distance and Algorithm Information". In: *Cognitive Research: Principles and Implications* 7.1, p. 70.
Partridge, Anthony and William B. Eldridge (1974). *The Second Circuit Sentencing Study: A Report to the Judges of the Second Circuit*. Tech. rep. Federal Judicial Center.
Pedersen, Kirstine Zinck and Anja Svejgaard Pors (2023). "Discretionary Responses in Frontline Encounters: Balancing Standardization with the Ethics of Office". In: *Journal of Public Administration Research and Theory* 33.1, pp. 80–93.
Pedersen, Mogens J. and Vibeke L. Nielsen (2020). "Bureaucratic Decision-Making: A Multi-Method Study of Gender Similarity Bias and Gender Stereotype Beliefs". In: *Public Administration* 98.2, pp. 424–40.
Pedersen, Mogens J. and Justin M. Stritch (2018). "RNICE model: Evaluating the Contribution of Replication Studies in Public Administration". In: *Public Administration Review* 78.4, pp. 606–12.
Pedersen, Mogens Jin, Justin M. Stritch, and Frederik Thuesen (2018). "Punishment on the Frontlines of Public Service Delivery: Client Ethnicity and Caseworker Sanctioning Decisions in a Scandinavian Welfare State". In: *Journal of Public Administration Research and Theory* 28.3, pp. 339–354.
Peeters, Rik and Arjan Widlak (2018). "The digital cage: Administrative exclusion through information architecture–The case of the Dutch civil registry's master data management system". In: *Government Information Quarterly* 35.2, pp. 175–183.
Peeters, Rik and Arjan C Widlak (2023). "Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal". In: *Public Administration Review* 83.4, pp. 863–877.

Pepin, Gabrielle, Christopher J. O'Leary, and Dallas Oberlee (2021). "Nudges to increase completion of welfare applications: Experimental evidence from Michigan". In: *Journal of Behavioral Public Administration* 4.2, pp. 1–8.
Peters, B. Guy (2018). *The Politics of Bureaucracy: An Introduction to Comparative Public Administration*. 7th. New York, NY: Routledge.
Pettit, Philip (1997). *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press.
Pfaff, Steven et al. (2021). "Do Street-Level Bureaucrats Discriminate Based on Religion? A Large-Scale Correspondence Experiment among American Public School Principals". In: *Public Administration Review* 81.2, pp. 244–259.
Pierre, Jon and B. Guy Peters (2020). *Governance, Politics and the State*. 2nd. London: Bloomsbury Academic.
Porumbescu, Gregory A., Vishal Trehan, and Agbonlahor Edomwonyi (2026). "Frontline Workers and Civic Tech: Bridging the Responsiveness Gap in Digital Client Encounters". In: *Public Administration Review* 86.4, pp. 1045–1059.
Prendergast, Canice (2007). "The Motivation and Bias of Bureaucrats". In: *American Economic Review* 97.1, pp. 180–196.
Pressman, Jeffrey L. and Aaron Wildavsky (1973). *Implementation: How Great Expectations in Washington Are Dashed in Oakland*. Berkeley: University of California Press.
Puranam, Phanish (2021). "Human–AI Collaborative Decision-Making as an Organization Design Problem". In: *Journal of Organization Design* 10.2, pp. 75–80.
Raaphorst, Nadine (2021). "Administrative Justice in Street-Level Decision-Making: Equal Treatment and Responsiveness". In: *The Oxford Handbook of Administrative Justice*. Ed. by Marc Hertogh et al. Oxford Handbooks, pp. 397–414.
Raaphorst, Nadine et al. (2026). "Equity and Equality in Street-Level Public Service Delivery". In: *Perspectives on Public Management and Governance* 9.1, pp. 7–19.
Rahwan, Iyad (2018). "Society-in-the-Loop: Programming the Algorithmic Social Contract". In: *Ethics and Information Technology* 20.1, pp. 5–14.
Raisch, Sebastian and Sebastian Krakowski (2021). "Artificial Intelligence and Management: The Automation–Augmentation Paradox". In: *Academy of Management Review* 46.1, pp. 192–210.
Ramji-Nogales, Jaya, Andrew I Schoenholtz, and Philip G Schrag (2007). "Refugee roulette: Disparities in asylum adjudication". In: *Stan. L. Rev.* 60, p. 295.
Rateb, George Botros Zarif (2022). "The Case for Clemency: Differential Impacts of Pretrial Detention on Case and Crime Outcomes". Honors thesis. Duke University.
Ravel, Katie et al. (2022). "Effect of Personalized Outreach on Medicaid to Marketplace Coverage Transitions". In: *JAMA Health Forum* 3.10, e223616.
Rehaag, Sean (2008). "Troubling Patterns in Canadian Refugee Adjudication". In: *Ottawa Law Review* 39.2, pp. 335–365.
— (2012). "Judicial Review of Refugee Determinations: The Luck of the Draw?" In: *Queen's Law Journal* 38.1, pp. 1–58.
— (2019). "Judicial Review of Refugee Determinations (II): Revisiting the Luck of the Draw". In: *Queen's Law Journal* 45.1, pp. 1–36.
— (2024). "Luck of the Draw III: Using AI to Extract Data About Decision-Making in Federal Court Stays of Removal". In: *Queen's Law Journal* 49.2, pp. 73–126.
Resh, William G. (2017). "Gary Miller and Andrew Whitford's 'Above Politics: Bureaucratic Discretion and Credible Commitment'". In: *Journal of Public Administration Research and Theory* 27.3, pp. 539–542.
Riccucci, Norma M. and Gregg G. Van Ryzin (2017). "Representative Bureaucracy: A Lever to Enhance Social Equity, Coproduction, and Democracy". In: *Public Administration Review* 77.1, pp. 21–30.
Robinson, Thomas S. and Raymond Duch (2024). "How to Detect Heterogeneity in Conjoint Experiments". In: *The Journal of Politics* 86.2, pp. 412–427.
Rogers, Parker (2025). "Retargeting Social Media Outreach for SNAP Take-Up". In: *Management Science* 72.6, pp. 5381–5387.
Rosenthal, Robert (1979). "The File Drawer Problem and Tolerance for Null Results". In: *Psychological Bulletin* 86.3, pp. 638–641.
Sackett, Paul R. et al. (2022). "Revisiting Meta-Analytic Estimates of Validity in Personnel Selection: Addressing Systematic Overcorrection for Restriction of Range". In: *Journal of Applied Psychology* 107.11, pp. 2040–68.
Sampat, Bhaven N. and Heidi L. Williams (2019). "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome". In: *American Economic Review* 109.1, pp. 203–236.
Schanzenbach, Diane (2009). *Experimental estimates of the barriers to Food Stamp enrollment*. Institute for Research on Poverty, University of Wisconsin-Madison Wisconsin.
Schauer, Frederick (1991). *Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*. Oxford: Clarendon Press.
Schiff, Daniel S., Kaylyn J. Schiff, and Patrick Pierson (2022). "Assessing Public Value Failure in Government Adoption of Artificial Intelligence". In: *Public Administration* 100.3, pp. 653–73.

Schillemans, Thomas and Marija Aleksovska (2025). "Making Nonpunitive Accountability Matter: Exploring Behavioral Effects of Nonpunitive Accountability in a Conjoint Experiment". In: *Public Administration*.
Schmidt, Eduard, Bernard Bernards, and Suzan van der Pas (2025). "Breaking the Rules, but for Whom? How Client Characteristics Affect Frontline Professionals' Prosocial Rule-Breaking Behavior". In: *Journal of Public Administration Research and Theory* 35.3, pp. 367–381.
Schmitz, Chris, Lewis Hammond, and Alan Chan (2026). *Characterizing Agentic Flooding of Government Services*. arXiv:2608.16603v2 [cs.CY]. Forthcoming, Proceedings of the 9th AAAI Conference on AI, Ethics, and Society (AIES).
Schooler, Jonathan (2011). "Unpublished results hide the decline effect". In: *Nature* 470.7335, pp. 437–437.
Schwalbach, Jan et al. (2025). *ParlLawSpeech*. Version 1.0.0. Data set.
Selin, Jennifer L (2019). "The best laid plans: How administrative burden complicates voting rights restoration law and policy". In: *Mo. L. Rev.* 84, p. 999.
Selten, Friso, Marcel Robeer, and Stephan Grimmelikhuijsen (2023). "'Just Like I Thought': Street-Level Bureaucrats Trust AI Recommendations If They Confirm Their Professional Judgment". In: *Public Administration Review* 83.2, pp. 263–278.
Senninger, Roman (Oct. 2023). "What makes policy complex?" In: *Political Science Research and Methods* 11.4, pp. 913–920.
Shayo, Moses and Asaf Zussman (2011). "Judicial Ingroup Bias in the Shadow of Terrorism". In: *Quarterly Journal of Economics* 126.3, pp. 1447–1484.
Shen, Menghan et al. (2024). "Migrant children's take-up of social health insurance: Experimental evidence from China". In: *Journal of Policy Analysis and Management*.
Shepard, Mark and Myles Wagner (Dec. 2022). *Do Ordeals Work for Selection Markets? Evidence from Health Insurance Auto-Enrollment*. Working Paper.
Shklar, Judith N. (1990). *The Faces of Injustice*. New Haven: Yale University Press.
Shoub, Kelsey (2022). "Comparing Systemic and Individual Sources of Racially Disparate Traffic Stop Outcomes". In: *Journal of Public Administration Research and Theory* 32.2, pp. 236–251.
Shumway, Clayson and Riley Wilson (2022). "Workplace Disruptions, Judge Caseloads, and Judge Decisions: Evidence from SSA Judicial Corps Retirements". In: *Journal of Public Economics* 205, p. 104573.
Sidhu, Dario et al. (2024). "Why Do Citizens Support Algorithmic Government?" In: *Journal of Public Policy* 44.3, pp. 659–677.
Sigstad, Henrik (2026). "Monotonicity among Judges: Evidence from Judicial Panels and Consequences for Judge IV Designs". In: *American Economic Review* 116.1, pp. 189–208.
Silver, Nate (2012). *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't*. New York: Penguin Press.
Simon, Herbert A. (1947). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. New York: Macmillan.
— (1997). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*. 4th ed. First edition 1947. New York: The Free Press.
Simonsohn, Uri (2007). "Clouds Make Nerds Look Good: Field Evidence of the Impact of Incidental Factors on Decision Making". In: *Journal of Behavioral Decision Making* 20.2, pp. 143–152.
Simonsohn, Uri, Joseph Simmons, and Leif D Nelson (2022). "Above averaging in literature reviews". In: *Nature Reviews Psychology* 1.10, pp. 551–552.
Skytte, Rasmus et al. (forthcoming). *Styringsmæssige rammer for den lokale autonomi*. Forskningsrapport til Magtudredningen 2.0. VIVE.
Slade, Patrick et al. (2024). "On Human-in-the-Loop Optimization of Human–Robot Interaction". In: *Nature* 633.8031, pp. 779–788.
Slough, Tara and Christopher J. Fariss (2021). "Misgovernance and Human Rights: The Case of Illegal Detention without Intent". In: *American Journal of Political Science* 65.1, pp. 148–165.
Slough, Tara and Scott A Tyson (2023). "External Validity and Meta-Analysis". In: *American Journal of Political Science* 67.2, pp. 440–455.
Soll, Jack B. and Richard P. Larrick (2009). "Strategies for Revising Judgment: How (and How Well) People Use Others' Opinions". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.3, pp. 780–805.
Sousa, Weslei G. de et al. (2019). "How and Where Is Artificial Intelligence in the Public Sector Going? A Literature Review and Research Agenda". In: *Government Information Quarterly* 36.4, Article 101392.
Spenkuch, Jörg L., Edoardo Teso, and Guo Xu (2023). "Ideology and Performance in Public Organizations". In: *Econometrica* 91.4, pp. 1171–1203.
St. John, Victor J. et al. (2026). "Administrative Burden in Higher Education: Race, Criminal Records, and Street-Level Bureaucrats in College Admissions". In: *Public Administration Review* 86.3, pp. 662–674.
Steijn, Bram and Joris van der Voet (2019). "Relational job characteristics and job satisfaction of public sector employees: When prosocial motivation and red tape collide". In: *Public Administration* 97.1, pp. 64–80.

Stenderup, Rasmus and Mogens Jin Pedersen (2025). "Compliance Burden versus Program Integrity Protection: A Survey Experiment on Citizen Attitudes toward Administrative Documentation Requirements in Public Service Policy". In: *Public Administration Review* 85.5, pp. 1512–1527.

Stensöta, Helena Olofsdotter (2012). "Political Influence on Street-Level Bureaucratic Outcome: Testing the Interaction between Bureaucratic Ideology and Local Community Political Orientation". In: *Journal of Public Administration Research and Theory* 22.3, pp. 553–571.

Stevenson, Megan T. and Jennifer L. Doleac (2024). "Algorithmic Risk Assessment in the Hands of Humans". In: *American Economic Journal: Economic Policy* 16.4, pp. 382–414.

Su, Min (2021). "Discretion in Traffic Stops: The Influence of Budget Cuts on Traffic Citations". In: *Public Administration Review* 81.3, pp. 446–458.

Suchman, Mark C. (1995). "Managing Legitimacy: Strategic and Institutional Approaches". In: *Academy of Management Review* 20.3, pp. 571–610.

Sunstein, Cass R. (2005). "Group Judgments: Statistical Means, Deliberation, and Information Markets". In: *New York University Law Review* 80.3, pp. 962–1049.

— (2022). "Governing by Algorithm? No Noise and (Potentially) Less Bias". In: *Duke Law Journal* 71.6, pp. 1175–1205.

Surowiecki, James (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday.

Svara, James H. (1998). "The Politics-Administration Dichotomy Model as Aberration". In: *Public Administration Review* 58.1, pp. 51–58.

Tetlock, Phillip E. and Dan Gardner (2015). *Superforecasting: The Art and Science of Prediction*. New York: Crown Publishers.

Tummers, Lars (2017). "The Relationship Between Coping and Job Performance". In: *Journal of Public Administration Research and Theory* 27.1, pp. 150–162.

Tummers, Lars and Victor Bekkers (2014). "Policy implementation, street-level bureaucracy, and the importance of discretion". In: *Public Management Review* 16.4, pp. 527–547.

Tummers, Lars and Philip Rocco (2015). "Serving Clients When the Server Crashes: How Frontline Workers Cope with E-Government Challenges". In: *Public Administration Review* 75.6, pp. 817–827.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157, pp. 1124–1131.

Tyler, Tom R. (2006). "Psychological Perspectives on Legitimacy and Legitimation". In: *Annual Review of Psychology* 57, pp. 375–400.

Tyler, Tom R. and Jonathan Jackson (2014). "Popular Legitimacy and the Exercise of Legal Authority: Motivating Compliance, Cooperation, and Engagement". In: *Psychology, Public Policy, and Law* 20.1, pp. 78–95.

Ulery, Bradford T. et al. (2012). "Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners". In: *PLoS ONE* 7.3, e32800.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone (2024). "When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis". In: *Nature Human Behaviour* 8, pp. 2293–2303.

Van de Walle, Steven and Geert Bouckaert (2003). "Public Service Performance and Trust in Government: The Problem of Causality". In: *International Journal of Public Administration* 26.8-9, pp. 891–913.

Van Gestel, RAF et al. (2023). "Improving take-up by reaching out to potential beneficiaries. Insights from a large-scale field experiment in Belgium". In: *Journal of Social Policy* 52.4, pp. 740–760.

Van Noordt, Colin and Gianluca Misuraca (2022). "Artificial Intelligence for the Public Sector: Results of Landscaping the Use of AI in Government Across the European Union". In: *Government Information Quarterly* 39.3, Article 101714.

Van Ryzin, Gregg G. (2011). "Outcomes, Process, and Trust of Civil Servants". In: *Journal of Public Administration Research and Theory* 21.4, pp. 745–60.

Vannoni, Matia, Elliott Ash, and Massimo Morelli (Jan. 2021). "Measuring Discretion and Delegation in Legislative Texts: Methods and Application to US States". In: *Political Analysis* 29.1, pp. 43–57.

Visser, E Lianne and Merlijn van Hulst (2024). "The Performance and Development of Deliberative Routines: A Practice-Based Ethnographic Study". In: *Journal of Public Administration Research and Theory* 34.1, pp. 92–104.

Visser, E. Lianne and Peter M. Kruyen (2021). "Discretion of the Future: Conceptualizing Everyday Acts of Collective Creativity at the Street-Level". In: *Public Administration Review* 81.4, pp. 676–690.

Waldron, Jeremy (2011). "The Rule of Law and the Importance of Procedure". In: *Getting to the Rule of Law: NOMOS L*. Ed. by James E. Fleming. New York: New York University Press, pp. 3–31.

Walker, Richard M. et al. (2019). "Best Practice Recommendations for Replicating Experiments in Public Administration". In: *Journal of Public Administration Research and Theory* 29.4, pp. 609–26.

Weber, Max (1978). *Economy and Society: An Outline of Interpretive Sociology*. Ed. by Guenther Roth and Claus Wittich. Originally published in German 1922. Berkeley: University of California Press.

Wei, Wenchi, Xin Chen, and Xing Ni (2026). "Awareness, Costs, and Take-Up of Social Benefits: Evidence from a Field Experiment in China". In: *Journal of Public Administration Research and Theory*. Forthcoming.

Wermink, Hilde T. et al. (2024). "Estimating Effects of Short-Term Imprisonment on Crime Using Random Judge Assignments". In: *Justice Quarterly* 41.3, pp. 317–346.
White, Ariel (2019). "Misdemeanor Disenfranchisement? The Demobilizing Effects of Brief Jail Spells on Potential Voters". In: *American Political Science Review* 113.2, pp. 311–324.
White, Ariel R., Noah L. Nathan, and Julie K. Faller (2015). "What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials". In: *American Political Science Review* 109.1, pp. 129–142.
White, Leonard D. (1926). *Introduction to the Study of Public Administration*. New York, NY: The Macmillan Company.
Widlak, Arjan and Rik Peeters (2020). "Administrative Errors and the Burden of Correction and Consequence: How Information Technology Exacerbates the Consequences of Bureaucratic Mistakes for Citizens". In: *International Journal of Electronic Governance* 12.1, pp. 40–56.
Williams, Jenny and Don Weatherburn (2022). "Can Electronic Monitoring Reduce Reoffending?" In: *Review of Economics and Statistics* 104.2, pp. 232–245.
Wilson, James Q. (1989). *Bureaucracy: What Government Agencies Do and Why They Do It*. New York: Basic Books.
Wilson, Woodrow (1887). "The Study of Administration". In: *Political Science Quarterly* 2.2, pp. 197–222.
Wirtz, Bernd W., Jan C. Weyerer, and Benjamin J. Sturm (2020). "The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration". In: *International Journal of Public Administration* 43.9, pp. 818–29.
Wooldridge, Jeffrey M. (2003). "Cluster-sample methods in applied econometrics". In: *American Economic Review* 93.2, pp. 133–138.
Wright, Bill J et al. (2017). "Low-cost behavioral nudges increase Medicaid take-up among eligible residents of Oregon". In: *Health Affairs* 36.5, pp. 838–845.
Wright, Bradley E. (2015). "The Science of Public Administration: Problems, Presumptions, Progress, and Possibilities". In: *Public Administration Review* 75.6, pp. 795–805.
Xu, Chengxin and Danbee Lee (2025). "No country for model minorities: Evidence of discrimination against Asian noncitizen immigrants in the U.S. nursing home market". In: *Public Administration Review* 85.5, pp. 1347–1364.
Yang, Crystal S. (2014). "Have Interjudge Sentencing Disparities Increased in an Advisory Guidelines Regime? Evidence from Booker". In: *New York University Law Review* 89.4, pp. 1268–1342.
Yin, Michelle, Garima Siwach, and Dajun Lin (2023). "Vocational Rehabilitation Services and Labor Market Outcomes for Transition-Age Youth with Disabilities in Maine". In: *Journal of Policy Analysis and Management* 42.1, pp. 166–197.
Yokum, David et al. (2022). "Effectiveness of Behaviorally Informed Letters on Health Insurance Marketplace Enrollment". In: *JAMA Health Forum* 3.3, e220034.
Young, Matthew M., Justin B. Bullock, and Jesse D. Lecy (2019). "Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration". In: *Perspectives on Public Management and Governance* 2.4, pp. 301–13.
Young, Matthew M. et al. (2024). "Complexity, Errors, and Administrative Burdens". In: *Public Management Review* 26.10, pp. 2847–2867.
Zacka, Bernardo (2017). *When the State Meets the Street: Public Service and Moral Agency*. Cambridge, MA: Harvard University Press.
Zhan, X., C. W.-H. Lo, and S.-Y. Tang (2014). "Contextual Changes and Environmental Policy Implementation: A Longitudinal Study of Street-Level Bureaucrats in Guangzhou, China". In: *Journal of Public Administration Research and Theory* 24.4, pp. 1005–1035.
Zhang, Yujie (2025). "Governance in layers: how female street-level bureaucrats reshape state power through emotional labor". In: *Journal of Public Administration Research and Theory* 35.3, pp. 323–337.
Zhirkov, Kirill (2022). "Estimating and Using Individual Marginal Component Effects from Conjoint Experiments". In: *Political Analysis* 30.2, pp. 236–249.

Danish summary

I en retsstat gælder det, at ens sager skal behandles ens. Hvordan borgerne bliver behandlet i mødet med staten, bør følge af loven og sagens fakta, ikke af hvem der forvalter den. Men hvorvidt bureaukratier lever op til dette ideal er et empirisk spørgsmål, da alle sager i sidste ende skal besluttes af bestemte bureaukrater, såsom dommere, politimænd, sagsbehandlere, læger, og jurister. Denne afhandling behandler på tværs af fire artikler det brede spørgsmål: hvor ofte, og hvorfor, får ens sager uens udfald?

Bureaukraters skøn står derfor centralt i afhandlingen. Tidligere forskning har fokuseret mest på, om bureaukrater i gennemsnit afviger fra, hvad en passende afgørelse måtte være; om systemet har bias; om den gennemsnitlige bureaukrat behandler forskellige borgere forskelligt. Der er derimod kun få studier, der har undersøgt, hvor meget bureaukrater afviger fra hinanden, hvis de skulle behandle den samme sag; om systemet har varians; om forskellige bureaukrater behandler lige sager forskelligt. Hvor det første spørgsmål holder bureaukrater konstant og ser på variansen i sager for så derefter at se på gennemsnit, holder det andet spørgsmål sagerne konstant og ser på variansen i sagsbehandlere for så at se på spredningen.

Denne afhandling behandler dette andet spørgsmål. Den ser altså på, i hvilket omfang afgørelser er vilkårlige. Kunne en given sag have fået et andet udfald, hvis den blot var blevet tildelt en anden bureaukrat? Det er et kontrafaktisk spørgsmål, og det kan måles uden nødvendigvis at skulle tjekke, hvad den korrekte afgørelse af en sag ville være.

For at måle på dette spørgsmål, kræver det, at hver bureaukrats bunke af sager i gennemsnit er den samme; derfor kræver det tilfældig tildeling af bureaukrater til sager. Når dette er tilfældet, kan man blot sammenligne forskellene i beslutningsrater mellem bureaukrater for at se, hvor mange som minimum ville have fået en anden beslutning, havde de fået en anden bureaukrat, da deres bunker i gennemsnit er ens. For at få adgang til sådanne data genanvender afhandlingen i dens første artikel de bagvedliggende data fra 95 leniency-instrumentvariabel-studier på tværs af forskellige typer af kontekster (for eksempel beslutninger om, hvorvidt børn skal tvangsfjernes, om folk skal varetægtsfængsles, eller om patenter skal godkendes). Dataet dækker mere end 25 millioner sager, afgjort af næsten 100.000 bureakrater.

I det mest normale tilfælde (mediankonteksten), afgør selve tildelingen af en bureaukrat mindst 13% af sagerne, og dette tal er mere end 29% i en fjerdedel af konteksterne. I de mest normale tilfælde (den midterste halvdel af konteksterne), vil to tilfældigt udtrukne bureaukrater være uenige i 3,3% til 12,9% af sagerne; hvis vi tror på, at en sag har én korrekt afgørelse, betyder uenighed, at den ene bureaukrat tog fejl, så i de fleste tilfælde er mindst 1,7% til 6,5% af afgørelserne i virkeligheden fejl.

Disse tal skyldes alene den forskel vi kan måle gennem bureaukratens gennemsnitlige eftergivenhed. Hvis der findes sager, hvor en i gennemsnit strengere bureaukrat vil være mere eftergiven end en i gennemsnit mere eftergiven bureaukrat, er tallene for små. Dette kunne være tilfældet, hvis der fx er forskelle i forskellige bureaukraters personlige bias mod forskellige karaktertræk. For at undersøge dette og holde eftergivenheden konstant, genanalyserer afhandlingen 13 conjoint-eksperimenter, hvor rigtige bureaukrater tvinges til at vælge mellem to forskellige

sager. Her er de som oftest uenige 40% af gangene.

Hvordan kan vi så mindske vilkårligheden og få ens sager til at blive behandlet ens? Afhandlingen trækker på to af sine artikler for at teoretisere over hvordan. Her peger afhandlingen på, at ved at tilbyde beslutningsstøtte til sagsbehandlerne, når de udøver skøn fx gennem AI og algoritmer, eller ved at gøre lovene og reglerne mere specifikke, kan man hjælpe sagsbehandlere til at blive mere konsistente.

Ens sager kan også få uens udfald, hvis borgeren ikke har den tilstrækkelige kunnen eller viden til at gennemføre ansøgningsprocesser til sociale ydelser og services. Den sidste artikel i afhandlingen slutter derfor med en metaanalyse af felteksperimenter, der søger at hjælpe borgere i disse situationer, og finder, at det på tværs af en lang række lande og forskellige typer ydelser er effektivt at give information til borgere og endnu mere effektivt at hjælpe dem gennem ansøgningsprocessen.







Appendix A

Reanalysis of the Israeli small-claims courts

This appendix shows the calculations behind the numbers comparing leniency and ethnicity bias using the data from Shayo and Zussman (2011). Many of the calculations are described rather briefly as they follow those of Paper 1.

Shayo and Zussman study small civil claims decided between 2000 and 2004; the outcome is whether the judge accepts the claim, and claims are assigned to judges as good as at random within a court. The conflict-period sample contains 1,748 cases and 132 judges. Many judges appear in very few cases, and a judge observed once carries no information about whether their apparent tendency would persist; the noise-corrected general component therefore uses the 98 judges observed at least twice and their 1,714 cases.

It is not a perfect setting for this kind of comparison since the number of cases per judge is still relatively low, but the point stands.

For the general leniency component, the estimand is the two-draw gap of equation 1 applied to persistent judge tendencies.

Raw judge means overstate the spread, as described in Paper 1. For this appendix we split each judge's cases into two random halves, estimate the tendency in each half, and keep only the covariance that persists across halves, using bootstrap. The corrected case-weighted variance is 0.00283 on the proportion scale, so $\widehat{\sigma}_\beta = 5.32$ pp. Under a normal reference, two judges drawn at random differ by $2\widehat{\sigma}_\beta/\sqrt{\pi} = 6.01$ pp, and this is a lottery only to the extent that the draw can land on someone else. Writing q_{ij} for the share of claimant i's court's cases heard by judge j, two draws return different judges with probability $\mathbb{E}_i[1 - \sum_j q_{ij}^2] = 0.68$, which gives $\mathrm{GMD}_\beta = 0.68 \times 6.01 = 4.10$ pp.

What does the ethnicity match do? The paper's full judge-fixed-effects specification reports a 19.23-percentage-point interaction between an Arab judge and an Arab claimant. This is a difference-in-differences on the observed claim-acceptance probability, so it is neither the effect of receiving a coethnic judge nor a variance across judges. From this we get: $\Delta_A = \Delta_J = \Delta$, which gives $\widehat{\Delta} = 19.23/2 = 9.61$ pp. This number assumes symmetry, but we could in theory estimate the two contrasts, one for Arabs and one for Jews, which gives 9.29 pp for Arab and 7.75 pp for Jewish claimants. The difference is not significant, however.

To get it on the same scale we want the GMD. Let H_{ij} indicate that claimant i and judge j are coethnic, and let m_i be the probability that the judge who receives claimant i's case is coethnic. Under the two-draw estimand, the Gini mean difference of the match channel alone is

$$\mathrm{GMD}_{\mathrm{match}} = \mathbb{E}_i\, \mathbb{E}_{j,k}[|\Delta\,(H_{ij} - H_{ik})|] = |\Delta|\, \mathbb{E}_i[2m_i(1 - m_i)]\,. \tag{25}$$

It is, however, only a lottery if it is even possible to change the ethnicity of the judge by chance. If every eligible judge is coethnic, the claimant always receives the premium and the

channel's GMD is zero.

The court files do not record the assignment probabilities, so we take the realised share of the claimant's court's cases heard by coethnic judges as a descriptive proxy for m_i. These shares vary heavily, however. In Haifa, an Arab-judge share of 0.43 gives an exposure $2m(1-m)$ of 0.49, while seventeen mostly Jewish courts have no Arab judges at all and thus there is an exposure of zero, the full-bias, zero-lottery.

Averaged over claims, we get $\mathbb{E}_i[2m_i(1-m_i)] = 0.2537$, and

$$\mathrm{GMD}_{\mathrm{match}} = |\widehat{\Delta}| \times \mathbb{E}_i[2m_i(1-m_i)] = 9.61 \times 0.2537 = 2.44 \text{ pp}, \tag{26}$$

We could calculate differences for Jews and Arabs, but they would be minor.

Now we have both on one scale, so the expected absolute change in the chance of winning when the judge is redrawn,

$$\mathrm{GMD}_{\beta} = 4.10 \text{ pp} \quad \text{against} \quad \mathrm{GMD}_{\mathrm{match}} = 2.44 \text{ pp}, \qquad \text{difference } 1.66 \text{ pp}, \quad \text{ratio } 1.68. \tag{27}$$

Importantly, the conditional premium itself is not small; when a redraw does change the match, the outcome moves by the full 9.61 points. The 2.44 is that premium times how rarely the draw flips the match. This was the whole idea behind the lottery component. It also shows a different thing, namely that the differences in leniency and the resulting arbitrariness occur all the time, whereas there are presumably fewer extremely discriminating officials.









Appendix B

Sources for Danish simplification initiatives

Table 19: Sources for Danish government initiatives to simplify regulation, 1982–2026. The table follows named national programmes and later high-profile sectoral initiatives identified in official sources; it does not include every ministry-level housekeeping project and does not measure whether the initiatives reduced regulation. Italics reproduce formal document titles, while bold text marks contemporary political labels when these were not document titles. The quotations reproduce the linked Danish-language sources without translation. Page references use the documents' printed pagination (*s.*); *sp.* refers to the column numbering in older volumes of *Folketingstidende*, for which the PDF page is also supplied. The 1982 slogan is documented in a later research paper hosted by the Danish Parliament because no paginated digital primary source was located. The 2008–09 row covers a process begun under Anders Fogh Rasmussen III and published as a formal plan under Lars Løkke Rasmussen I.

Year	Government	Initiative	Quotation
1982	Poul Schlüter I	**Political label:** Operation regelsanering. **Contemporary document title:** *Regelforenklingsarbejdet 1982–83 – ministeriernes redegørelser for regelforenklingsarbejdet inden for deres områder*	"Det skal være lettere at være dansker!" (s. 50; secondary source)
1988	Poul Schlüter III	*Handlingsplan for afbureaukratisering*	"Formålet med planen var at forenkle, effektivisere og afbureaukratisere den offentlige sektor." (s. 15)
1991	Poul Schlüter IV	**Ministerial label:** Operation Regelstorm	"Arbejdet med regelforenkling og regelfrihed vil fortsat være et område, som jeg prioriterer meget højt." (sp. 8287–8288; PDF-s. 14)
1995	Poul Nyrup Rasmussen II	**Political label:** "Kampen mod bøvlet". **Report title:** *Lettelse af administrative byrder – rapport fra udvalget om lettelse af administrative byrder for nye, små og mellemstore virksomheder*	"Det er regeringens agt at fortsætte arbejdet, forstærke indsatsen for at lette de administrative byrder." (sp. 6431; PDF-s. 4)

Continued on next page

Table 19 – continued

Year	Government	Initiative	Quotation
2002	Anders Fogh Rasmussen I	**Political label:** Kampagne mod Systemdanmark. **Action plan:** *En enklere offentlig sektor – Handlingsplan for regelforenkling og administrative lettelser*	"[. . .] lette administrative byrder for borgere og virksomheder og fjerne unødige regler i den offentlige sektor." (sp. 4434; PDF-s. 12)
2008–09	Anders Fogh Rasmussen III; Lars Løkke Rasmussen I	*Mere tid til velfærd – Regeringens plan for mindre bureaukrati i kommuner og regioner* (preparatory work in 2008; plan published 6 October 2009)	"I alt vil forenklingen af de statslige regler frigøre knap 900 millioner kr. årligt fra 2011 [. . .]" (s. 8)
2010	Lars Løkke Rasmussen I	*Enkel administration i staten* **Heading:** *Afbureaukratisering i stat, kommuner og regioner*	"[. . .] 30 forenklingsinitiativer, der giver færre og bedre regler, mere smidig sagsbehandling og klar kommunikation [. . .]" (s. 10)
2011	Lars Løkke Rasmussen I	*Væk med bøvlet – Et bedre møde mellem borgerne og det offentlige*	"Det er regeringens mål, at det skal være mindre bøvlet at være dansker." (s. 1)
2012	Helle Thorning-Schmidt I	*Enklere regler – den virksomhedsrettede indsats for bedre regulering 2012–2015* Including *Virksomhedsforum for enklere regler*	"De administrative byrder skal samlet set være lavere i 2015 end i 2011." (s. 82)
2013	Helle Thorning-Schmidt I	*Principper for samarbejde mellem parter på det offentlige arbejdsmarked om modernisering* **Heading:** *5.5 Bedre regulering og målretning*	"Det er målet at afskaffe opgaver og regler, der tager tid fra de offentlige kerneopgaver [. . .]" (s. 57)
2015	Lars Løkke Rasmussen II	*Sammen for fremtiden* **Heading:** *Bedre vilkår for virksomheder i Danmark*	"Regeringen vil herudover lette erhvervslivets administrative byrder med op mod 3 mia. kr. frem mod 2020." (s. 11)
2017	Lars Løkke Rasmussen III	*Enkle regler, mindre bureaukrati – lovgivning i en digital virkelighed*	"Der skal luges ud i unødvendige og komplekse regler [. . .]" (s. 3)
2017–18	Lars Løkke Rasmussen III	*Sammenhængsreformen – Mere frihed, tillid og tryghed* Including the campaign *Meld en regel* and the heading *Mindre topstyring og større lokal frihed*	"Staten, kommuner og regioner skal afskaffe unødvendige regler og registreringskrav." (s. 13)

Continued on next page

Table 19 – continued

Year	Government	Initiative	Quotation
2018	Lars Løkke Rasmussen III	*Færre regler og mindre bureaukrati*	"Vi skal reducere de regler og registreringskrav, som giver bøvl for borgerne og unødigt begrænser medarbejdernes muligheder for at yde borgernær velfærd." (s. 7)
2019	Lars Løkke Rasmussen III	*Mindre bureaukrati – mere vækst* **Subtitle:** *Udspil til at lette snærende regler og administrative byrder for virksomhederne*	"Virksomheder skal kunne fokusere på deres forretning i stedet for at skulle bruge tiden på at udfylde skemaer og navigere efter unødigt komplicerede regler." (s. 5)
2019	Lars Løkke Rasmussen III	*Enkel administration i staten* **Heading:** *En mere enkel administration i staten*	"29 initiativer, der styrker kvaliteten og skaber en mere enkel og værdiskabende statslig administration." (s. 5)
2019	Mette Frederiksen I	**Nærhedsreform** (announced in the prime minister's opening address)	"For vi skal jo ikke bare lovgive forfra, vi skal faktisk give slip [. . .]" (s. 4)
2020	Mette Frederiksen I	*Aftale om velfærdsaftaler på ældreområdet*	"Aftalen skal derfor sikre mulighed for, at ledere og medarbejdere kan frigøre sig mest muligt fra statslig og kommunal regulering [. . .]" (s. 1)
2022	Mette Frederiksen I	*Danmark kan mere III – Bedre uddannelser. Mere velfærd* **Heading:** *Regelstop*	"Regeringen foreslår et statsligt regelstop, som skal bremse væksten af nye regler og krav." (s. 11)
2022	Mette Frederiksen II	*Ansvar for Danmark – Det politiske grundlag for Danmarks regering* **Subheading:** *3.2 Frisættelse*	"Til gengæld afskaffes al unødig statslig og kommunal regulering." (s. 15–16)
2023	Mette Frederiksen II	*DK2030 – Danmark rustet til fremtiden* **Heading:** *Frisættelse og andre veje til bedre velfærd for pengene*	"Unødig statslig og kommunal regulering skal afskaffes." (s. 33)
2024	Mette Frederiksen II	*Hvorfor skal vi af med bøvl og bureaukrati for foreninger?*	"Det vil regeringen ændre på. Vi vil gøre op med de mange regler, der spænder ben for frivillige landet over." (s. 1)
2025	Mette Frederiksen II	*Aftale om reform af beskæftigelsesindsatsen – Mere værdighed, større frihed og færre regler*	"Beskæftigelsesreformen er den største forenkling og afbureaukratisering af beskæftigelsesindsatsen nogensinde." (s. 2)

Continued on next page

Table 19 – continued

Year	Government	Initiative	Quotation
2025	Mette Frederiksen II	*Flerårigt arbejdsprogram for staten* **Heading:** *Færre statslige regler og krav*	"Regeringen vil fortsætte arbejdet med at sanere eksisterende regler og krav og vil konsekvent afskaffe al unødig regulering [...]" (s. 10)
2026	Mette Frederiksen III	*Det politiske grundlag for firkløverregeringen* **Heading:** *Et stærkt erhvervsliv* **Subheading:** *Regeringens initiativer*	"Ekspertgruppen skal udpege regler, der er uhensigtsmæssigt byrdefulde, og på den baggrund introduceres et nyt 'afvikl eller forklar'-princip [...]" (s. 47)







Appendix C

External applications of the SEWP

Study	Setting	N	Initial SD	Revised SD	SEWP (95% CI)
Pálfi et al. (2022)	General practitioners estimated cancer risk for 20 shared patient vignettes before and after QCancer advice.	157	15.14	9.99	34.0% (27.5–41.5)
Fogliato et al. (2021)	Crowdworkers predicted rearrest risk before and after seeing a model prediction; 6,775 judgments.	271	0.201	0.177	12.0% (9.5–15.0)

Table 20: The two external SEWP applications closest to public-decision settings. Dispersion is computed within vignette or offender profile; intervals use 1,000 participant-clustered bootstrap draws. The scales differ across studies. Neither study observes officials making binding public decisions.









Appendix D

Paper 1 Appendices

D.1 Complier share among recipients: the P10–P90 contrast

Figure 36 in the appendix measures the complier share among rarer-outcome recipients under a one-standard-deviation shift in examiner tendency. Figure 29 repeats the exercise with the wider P10–P90 contrast, the same range used for the complier share in Figure 33. The shape is the same and the levels are higher: below a 5% base rate the median share of rarer-outcome recipients who are marginal is about 80%, against about 40% under the one-standard-deviation contrast.

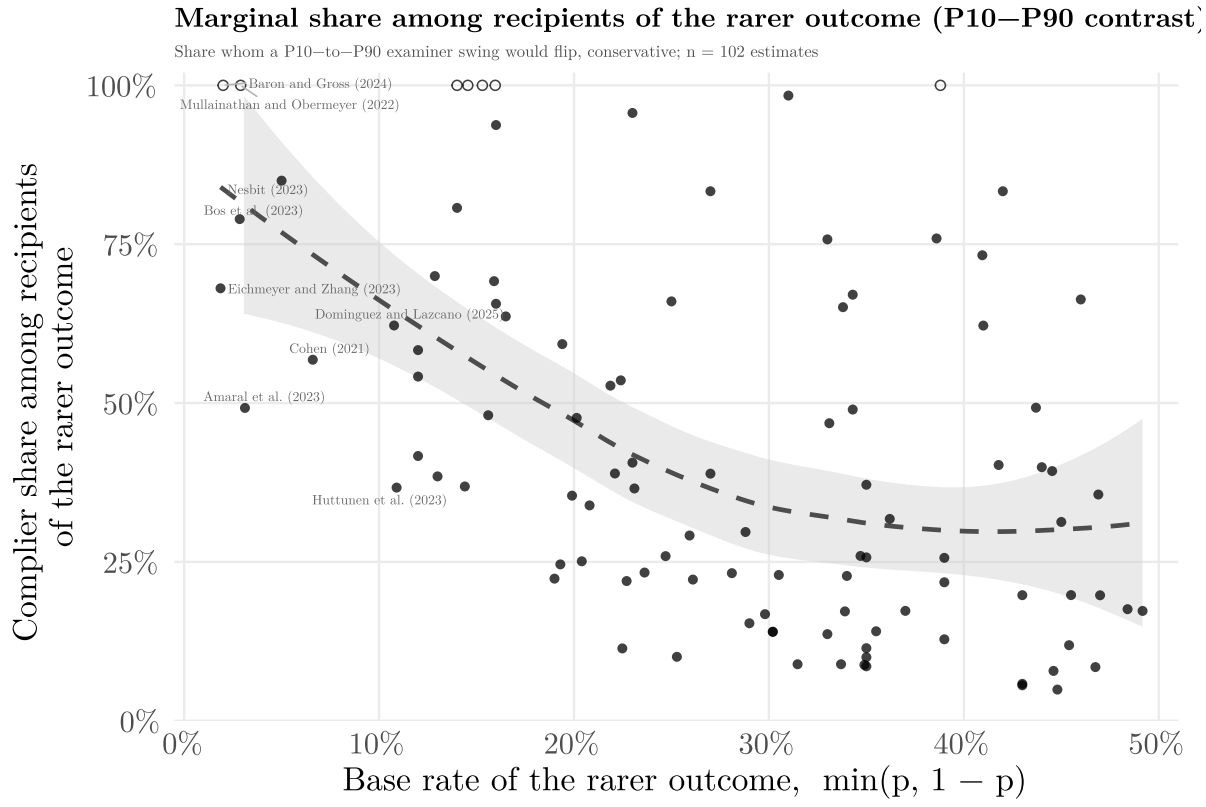


Figure 29: Complier share among recipients of the rarer outcome under the P10–P90 examiner contrast (companion to Figure 36). Half the P10–P90 range is divided by the rarer outcome's base rate $\min(p, 1-p)$ and capped at one. The share rises as the outcome becomes rarer (Spearman $\rho = -0.50$, $n = 102$). Hollow points mark studies where half the range exceeds the base rate, so essentially all recipients are marginal (7 studies). The P10–P90 range is conservative relative to min-max but wider than the one-standard-deviation contrast.

D.2 How the variance decomposition follows from the model

The model for the latent utility is

$$y^*_{ir} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_i'\mathbf{u}_r + \varepsilon_{ir},$$

where the differenced attribute columns $\mathbf{x}_i$ are set by the randomization, the personal deviations $\mathbf{u}_r \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$ with $\boldsymbol{\Sigma}_u = \mathrm{diag}(\sigma^2_{u_1}, \ldots, \sigma^2_{u_K})$ are drawn independently of $\mathbf{x}_i$, and the errors $\varepsilon_{ir} \sim N(0,1)$ are drawn independently of both. The variance I decompose is the variance of y^*_{ir} over the joint draw of a comparison, a bureaucrat, and an error.

$\boldsymbol{\beta}$ is fixed and $\mathbf{x}$ is random, so

$$\mathrm{Var}(\mathbf{x}'\boldsymbol{\beta}) = \mathrm{Cov}\Big(\sum_k \beta_k x_k,\ \sum_l \beta_l x_l\Big) = \sum_k \sum_l \beta_k \beta_l \,\mathrm{Cov}(x_k, x_l) = \boldsymbol{\beta}'\,\mathrm{Cov}(\mathbf{x})\,\boldsymbol{\beta}.$$

Since $\mathbf{u}_r$ is independent of $\mathbf{x}$ and has mean zero, $\mathbb{E}[\mathbf{x}'\mathbf{u}_r] = 0$, and therefore

$$\mathrm{Var}(\mathbf{x}'\mathbf{u}_r) = \mathbb{E}\big[(\mathbf{x}'\mathbf{u}_r)^2\big] = \mathbb{E}\big[\mathbf{u}_r'\,\mathbf{x}\mathbf{x}'\,\mathbf{u}_r\big] = \mathrm{tr}\big(\mathbb{E}[\mathbf{x}\mathbf{x}']\,\boldsymbol{\Sigma}_u\big).$$

With $\boldsymbol{\Sigma}_u$ diagonal, only the diagonal of $\mathbb{E}[\mathbf{x}\mathbf{x}']$ survives the trace:

$$\mathrm{Var}(\mathbf{x}'\mathbf{u}_r) = \sum_k \sigma^2_{u_k}\,\mathbb{E}[x_k^2].$$

The probit normalization fixes $\mathrm{Var}(\varepsilon) = 1$.

D.3 Variance decomposition model specification

I estimate a random-slopes probit model using Stan (Carpenter et al., 2017) via `cmdstanr`. The full model code is reproduced below.

The priors for $\boldsymbol{\beta}$ are $N(0,1)$ and for $\boldsymbol{\sigma}_u$ are half-normal$(0,1)$, chosen to be weakly informative on the probit scale. I use a non-centered parameterization ($\mathbf{u}_r = \mathbf{z}_r \odot \boldsymbol{\sigma}_u$, where $\mathbf{z}_r \sim N(\mathbf{0}, \mathbf{I})$) to improve sampling efficiency. The model is estimated with 4 chains, 1000 warmup and 1000 sampling iterations, with `adapt_delta = 0.95` and `max_treedepth = 12`.

Listing 1: Stan model for variance decomposition

```
data {
  int<lower=1> N;                          // total observations
  int<lower=1> I;                          // number of respondents
  int<lower=1> K;                          // number of attributes
  array[N] int<lower=1, upper=I> id;       // respondent index
  matrix[N, K] X;                          // attribute matrix (differenced)
  array[N] int<lower=0, upper=1> Y;        // binary choice
}
parameters {
  vector[K] beta;                          // fixed effects (agreement)
  vector<lower=0>[K] sigma_u;              // heterogeneity SDs
```

```
  matrix[I, K] z_u;                          // standardized random effects
}
transformed parameters {
  matrix[I, K] u = z_u .* rep_matrix(sigma_u', I);
}
model {
  beta ~ normal(0, 1);
  sigma_u ~ normal(0, 1);
  to_vector(z_u) ~ std_normal();
  vector[N] eta;
  for (n in 1:N) {
    eta[n] = dot_product(X[n], beta + u[id[n]]');
  }
  Y ~ bernoulli(Phi_approx(eta));
}
```

D.4 AMCE validation

To verify that the agreement component $\boldsymbol{\beta}$ in my probit model recovers the same quantities as standard conjoint analyses, I compare the probit marginal effects ($\beta_k \times \phi(0) \approx \beta_k \times 0.399$) with OLS-estimated AMCEs (a linear probability model without intercept on the differenced attributes). Table 21 reports the Pearson correlation between the two sets of estimates for each study. Across all studies with available data, the mean correlation is 0.998, confirming that my agreement component captures the same ranking of attributes as standard conjoint AMCEs. Because $\phi(0)$ is a constant scaling, this correlation tests the ranking and sign pattern of the attributes rather than the magnitude of the marginal effects.

As a further check, I compared my probit estimates against the AMCEs reported in the original papers. For Liu (2019), who reports OLS AMCEs for binary attributes, the ranking and signs of my probit marginal effects match exactly: elite university (paper: 0.115, mine: 0.133), CCP membership (0.079 vs. 0.081), male (0.065 vs. 0.084), and postgraduate degree (0.049 vs. 0.062). For Aleksovska et al. (2022), who report marginal means rather than AMCEs, the implied effect sizes also align: expertise (0.06 vs. 0.062) and previous experience (0.06 vs. 0.077). For Adam et al. (2021), who use a Bayesian logistic model and report odds ratios, the sign ordering matches my estimates: language proficiency is the dominant attribute, followed by nationality, with gender and age playing smaller roles.

Table 21: Correlation between OLS AMCEs and probit marginal effects across studies.

Study	Attributes (K)	r(OLS, Probit)
Aleksovska et al. (2022)	5	0.998
Schillemans and Aleksovska (2025)	5	1.000
Jankowski et al. (2020)	25	0.999
Liu (2019)	7	0.990
Liu (2024)	7	1.000
Dietrich et al. (2026)	14	0.997
Adam et al. (2021)	6	1.000
Fischer and Weißmüller (2024) (reporting)	4	1.000
Fischer and Weißmüller (2024) (correcting)	4	0.997
Mean		0.998

D.5 Disagreement and partisanship in Liu (2024)

Liu (2024) has the highest pairwise agreement (74.1%) among studies with shared comparisons. To check whether the remaining disagreements are driven by the partisan attribute, I compare agreement rates conditional on whether the two profiles in a comparison differ on partisanship. Agreement is somewhat higher when partisanship is held constant (77.8%) than when it differs (70.2%), a gap of 7.6 percentage points (bootstrap 95% CI [1.9, 13.7]). Holding partisanship fixed therefore raises agreement, though only modestly, and partisanship is not the dominant fault line. Among the rater-pairs who disagree, partisanship differs in 56.0% of cases, less often than resource allocation (74.2%), job-creation outcomes (64.9%), or collaborative trust (68.7%). Partisanship is thus the attribute that least often distinguishes the cases people disagree about, so disagreement is spread across attributes rather than concentrated on partisan divisions.

D.6 Geographic distribution of studies

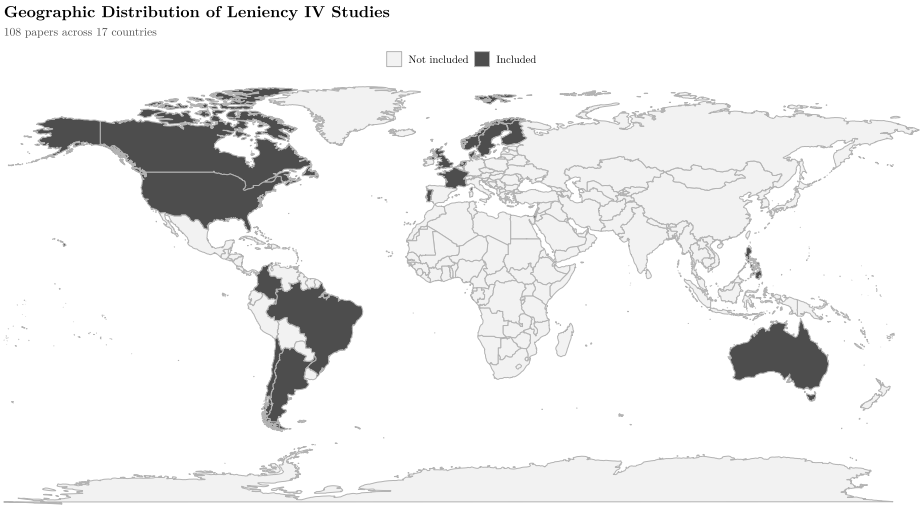


Figure 30: Geographic distribution of leniency IV studies. The map spans 108 studies across 17 countries, heavily concentrated in the United States (68 papers).

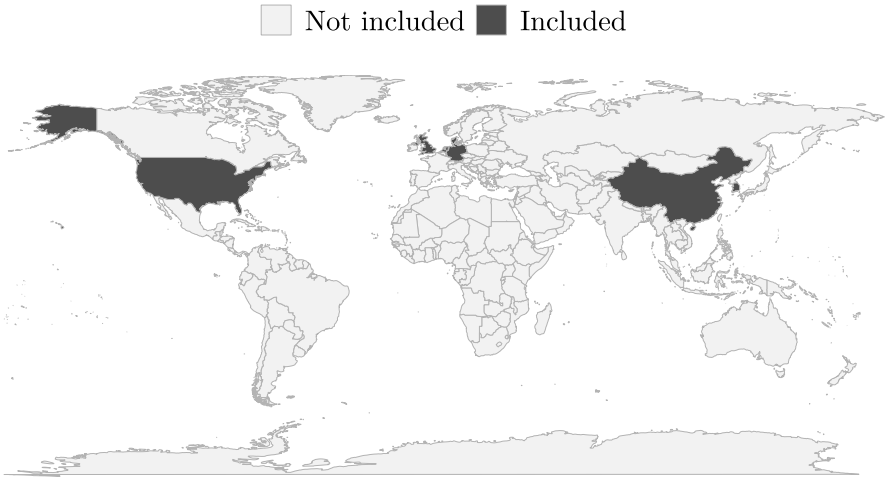


Figure 31: Geographic distribution of conjoint studies. The sample spans 13 studies across 7 countries, with Germany accounting for nearly half of all studies.

D.7 Additional leniency-IV results

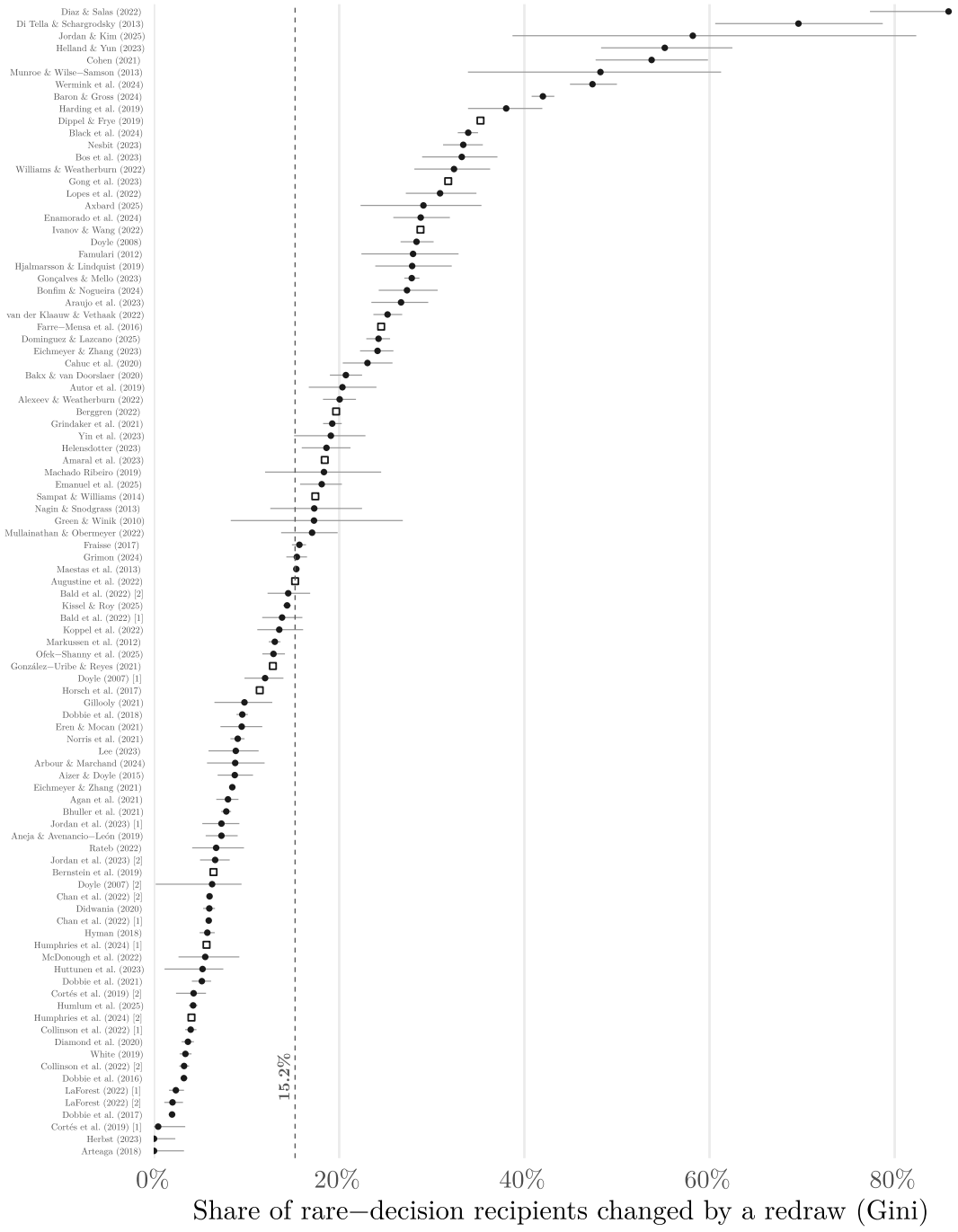


Figure 32: Of those who received the rarer decision, the share whose outcome a redraw of the bureaucrat would change, by study (the 95 estimates in the analysis sample). Computed as the corrected Gini mean difference divided by twice the rare-side base rate; conventions as in Figure 10: whiskers are 95% intervals, open squares mark estimates without such an interval, and the dashed line marks the median of 14.4%. The largest value (Diaz and Salas 2022) is capped at the feasibility bound of its base rate.

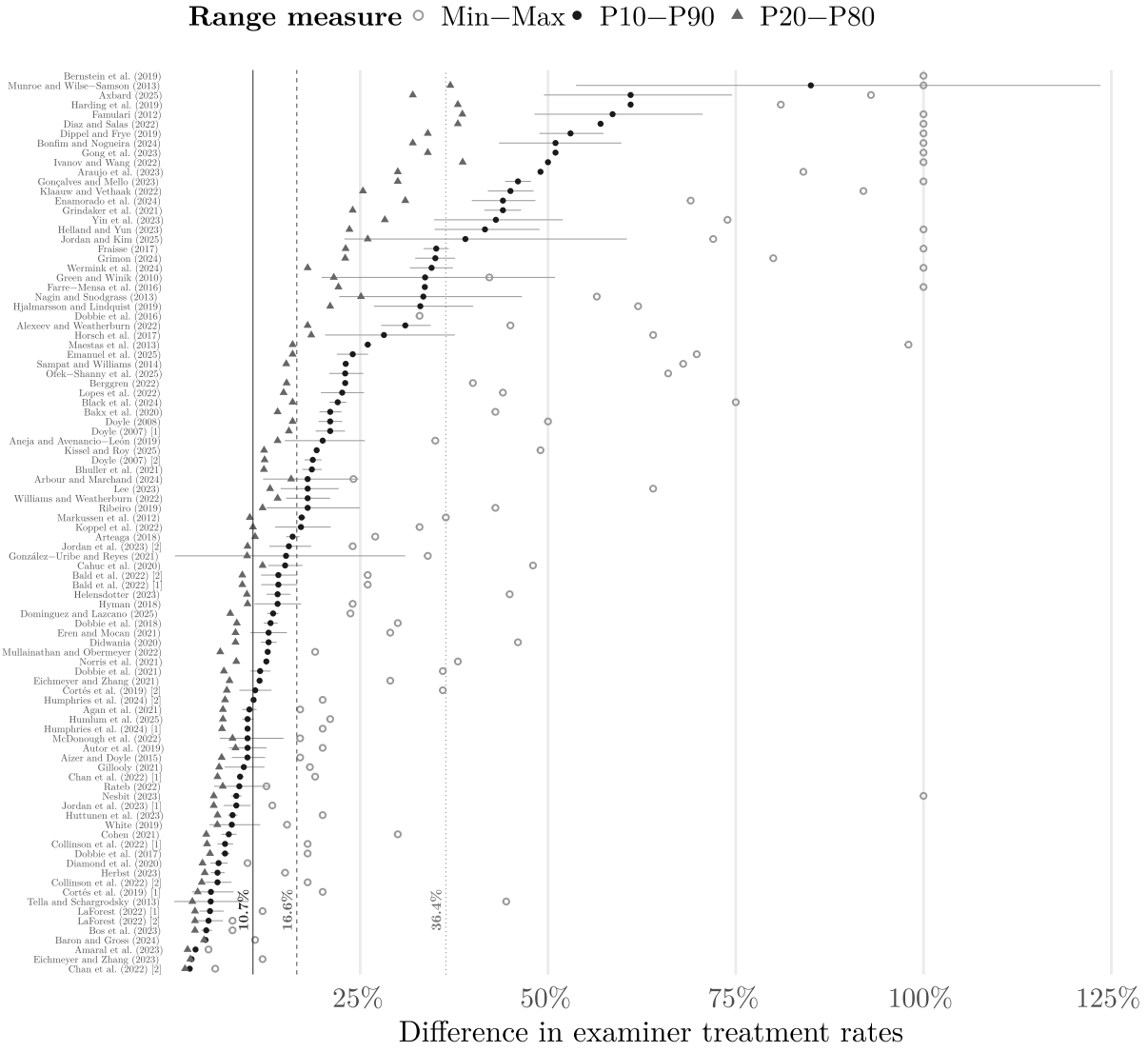


Figure 33: Range of examiner propensities across the 95 estimates in the analysis sample. Under a monotonicity assumption, the min-max range identifies the complier share, the fraction of cases whose outcome depends on examiner assignment. The p_{10}–p_{90} range is a conservative lower-bound proxy that excludes the tails of the examiner distribution. For studies using residualized instruments, min-max ranges exceeding one are capped at one; no p_{10}–p_{90} ranges exceed one. Whiskers give simulated 95% confidence intervals for the p_{10}–p_{90} range, built from each study's number of examiners, caseload, base rate, and examiner-propensity SD; they cover both the draw of examiners and per-examiner estimation error. Intervals for raw-propensity studies are truncated at one. Min-max and p_{20}–p_{80} are shown as points only, since confidence intervals for the extreme order statistics would rest on pure tail assumptions.

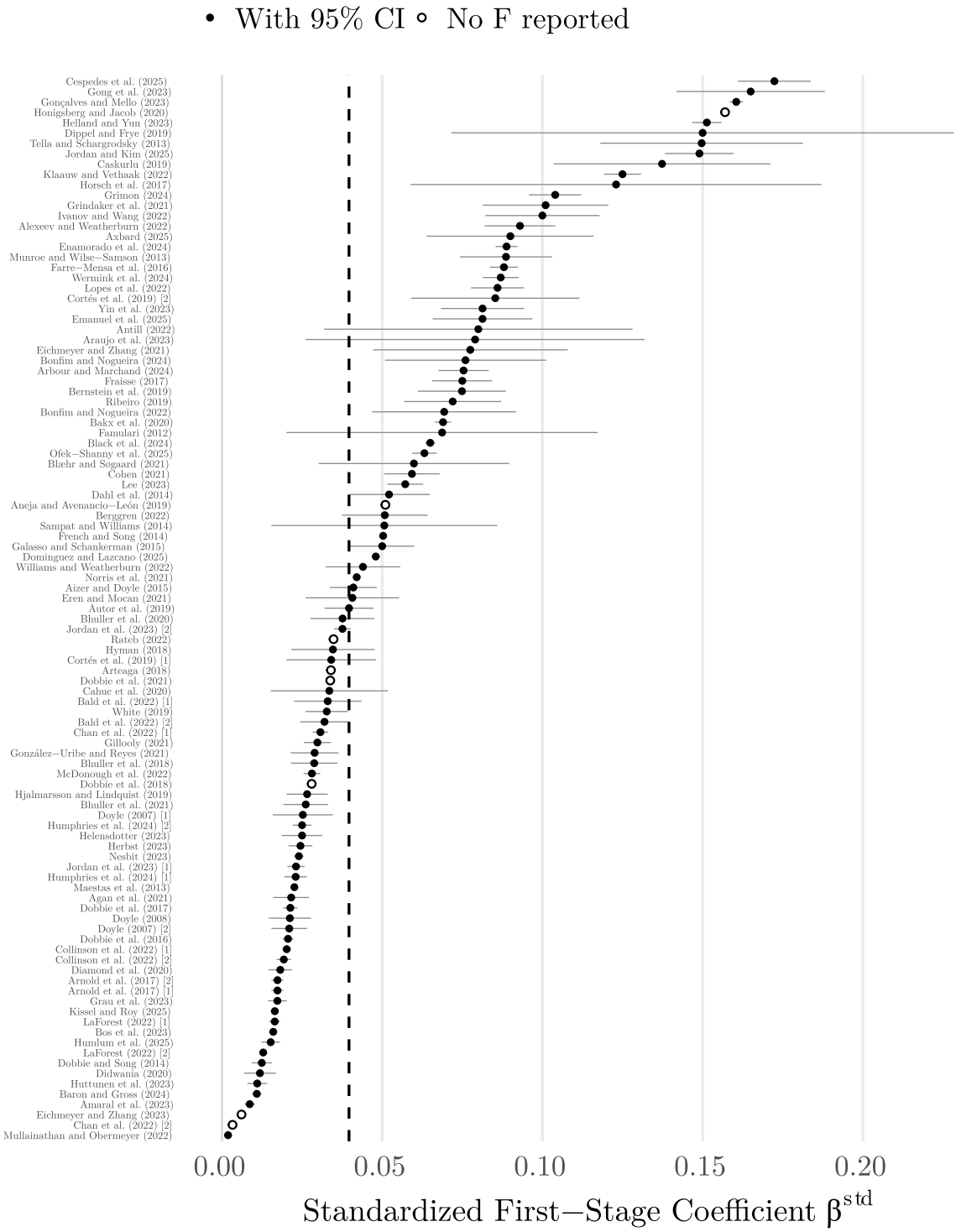


Figure 34: Standardized first-stage coefficients (β^{std}). Each point shows how much a one-standard-deviation increase in examiner leniency shifts the probability of the bureaucratic decision. Whiskers are 95% confidence intervals with $\text{SE} = |\beta^{\text{std}}|/\sqrt{F}$ derived from the reported first-stage F-statistic, which is exact when the reported F is that of the single leniency instrument (98 of 103 estimates); open circles mark estimates with no reported F. Because the F-statistics are as reported by the original authors, any inflation from improper clustering carries into the intervals, making them lower bounds on width in those cases.

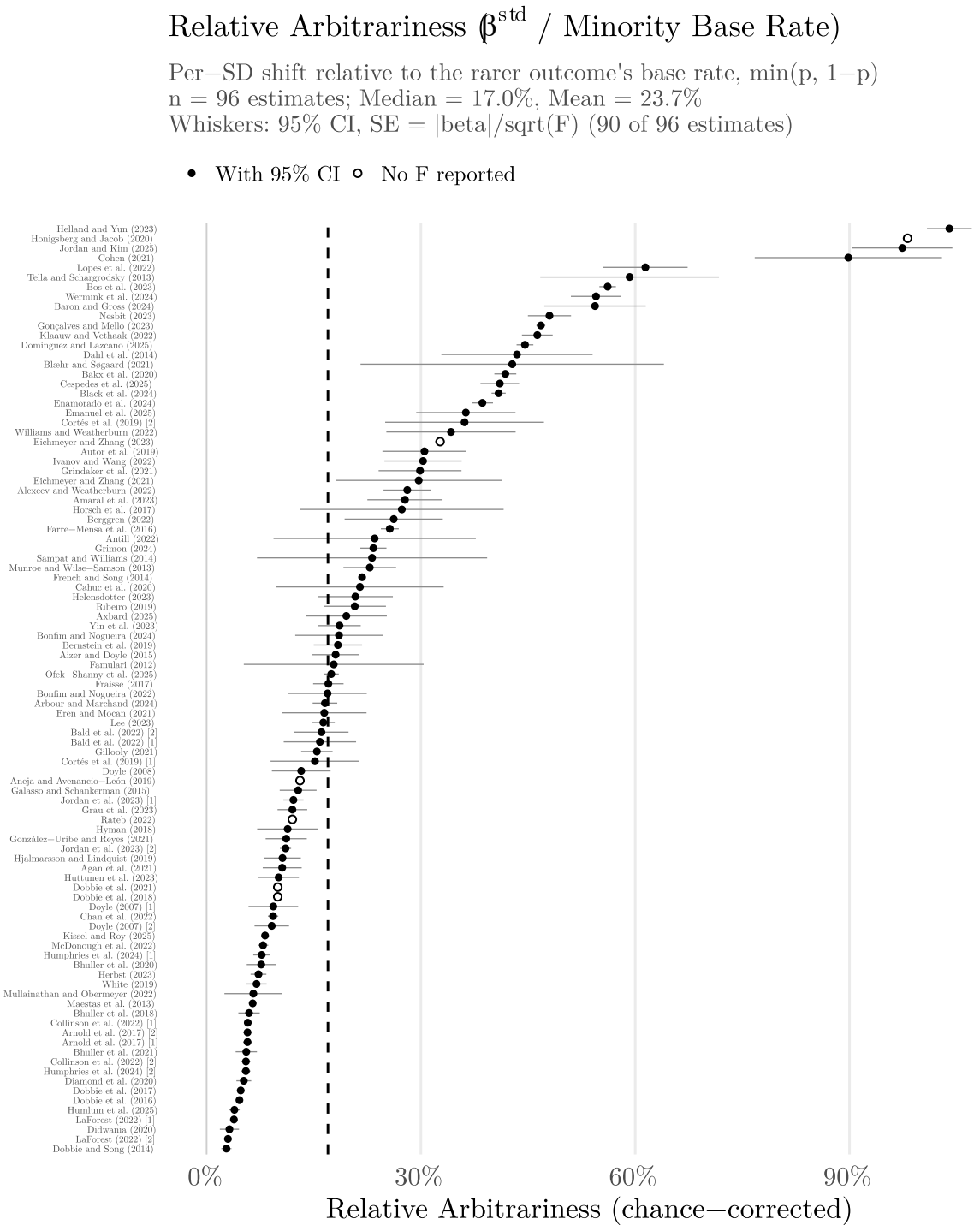


Figure 35: Relative arbitrariness (β^{std} divided by the lower of the two base rates, $\min(p, 1-p)$, the chance-corrected measure). This weights the standardized coefficient by the rarer outcome's baseline probability, capturing the proportional impact of bureaucrat assignment. Whiskers are 95% confidence intervals: the β^{std} interval, with SE $= |\beta^{\text{std}}|/\sqrt{F}$ from the reported first-stage F-statistic, divided by the same base rate (92 of 96 estimates); open circles mark estimates with no reported F.

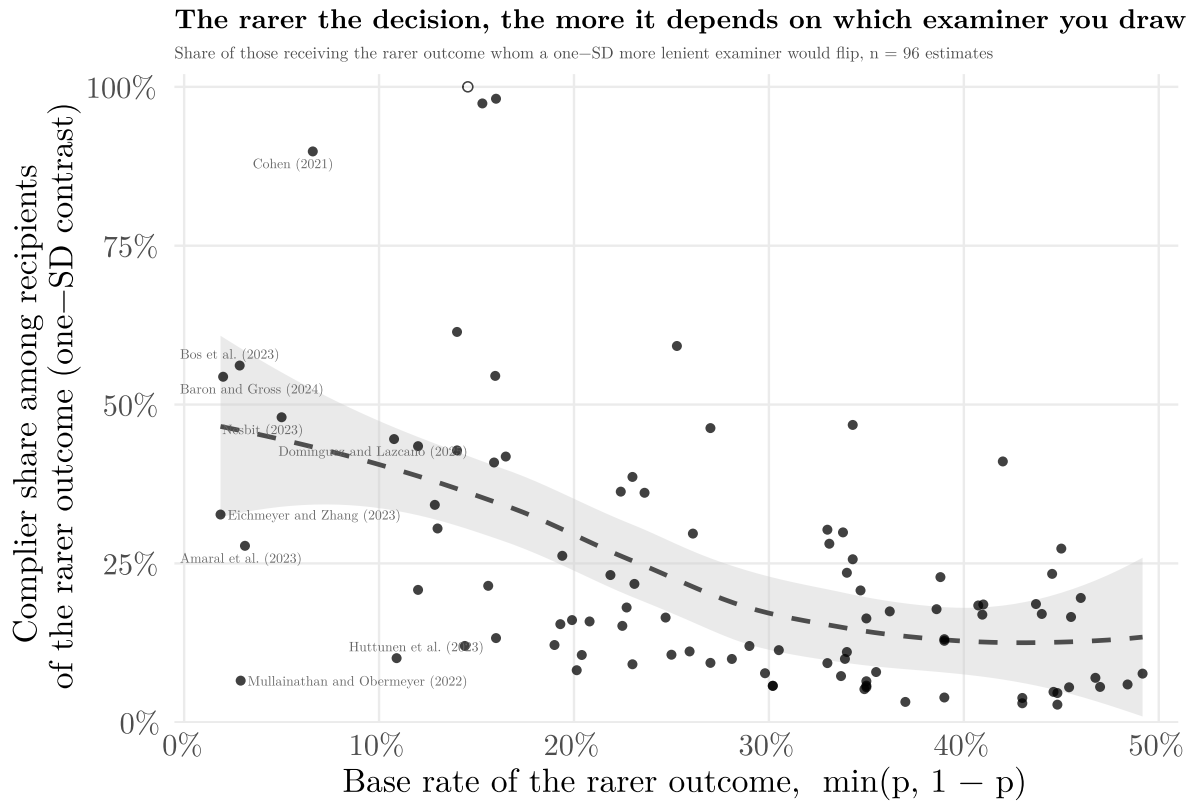


Figure 36: Complier share among recipients of the rarer outcome under a one-standard-deviation shift in examiner tendency.

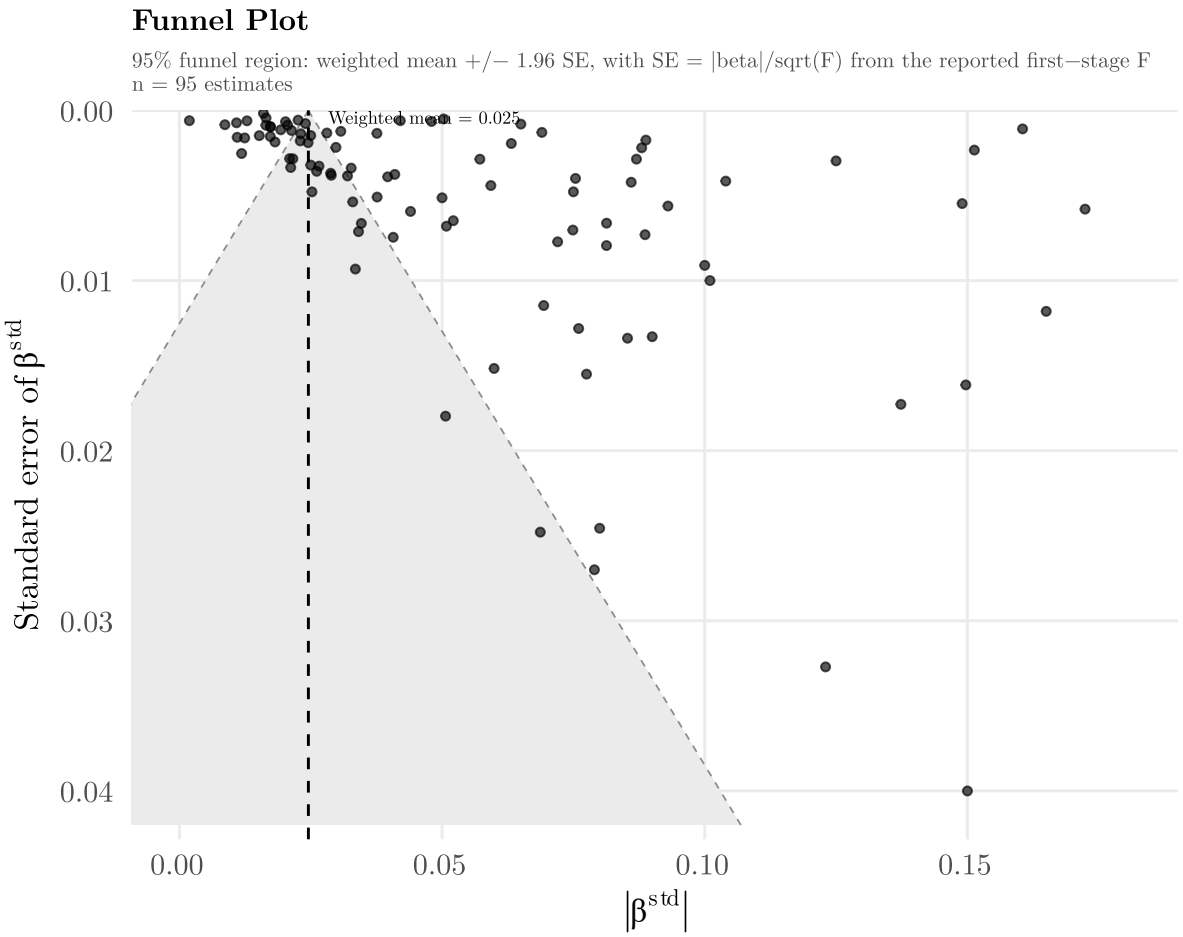


Figure 37: Funnel plot of standardized first-stage coefficients ($|\beta^{\text{std}}|$) against their standard errors, with SE = $|\beta^{\text{std}}|/\sqrt{F}$ from the reported first-stage F-statistic; more precise estimates sit higher. The grey wedge is the 95% funnel region around the inverse-variance-weighted mean (dashed line). Estimates at every precision level lie far outside the wedge, so the cross-study spread reflects genuine heterogeneity in examiner dispersion across settings rather than sampling noise around a common value. No Egger test is reported: with the standard error derived from β and F, the Egger regression is mechanically degenerate.

D.8 IV assumptions in the underlying studies

The exclusion restriction requires that examiner assignment affects outcomes only through the treatment decision. Two-thirds of studies acknowledge that examiners could plausibly affect outcomes through other channels, since a judge might influence not just the incarceration decision but also sentence length, bail conditions, or courtroom interactions, and about half formally test whether the instrument predicts other treatment margins. Monotonicity requires that a stricter examiner is stricter for all cases, not just on average; 69% of studies test this, typically by showing positive first-stage coefficients across subgroups. Passing carries little weight: Sigstad (2026) applies the tests used in this literature to judicial-panel data, where several judges vote on the same case and violations are directly visible, and the tests detect none of them. These rates therefore document the literature's practice rather than establish the assumption; as argued in Study 1, violations would make my reported ranges understate arbitrariness rather than

overstate it.

Instrument strength is summarized in Figure 38. Of the 109 main estimates reporting a first-stage F-statistic, the median is 120.6 and the interquartile range is 53.5–448.2. Seven estimates (6%) fall below the conventional weak-instrument threshold of $F = 10$.

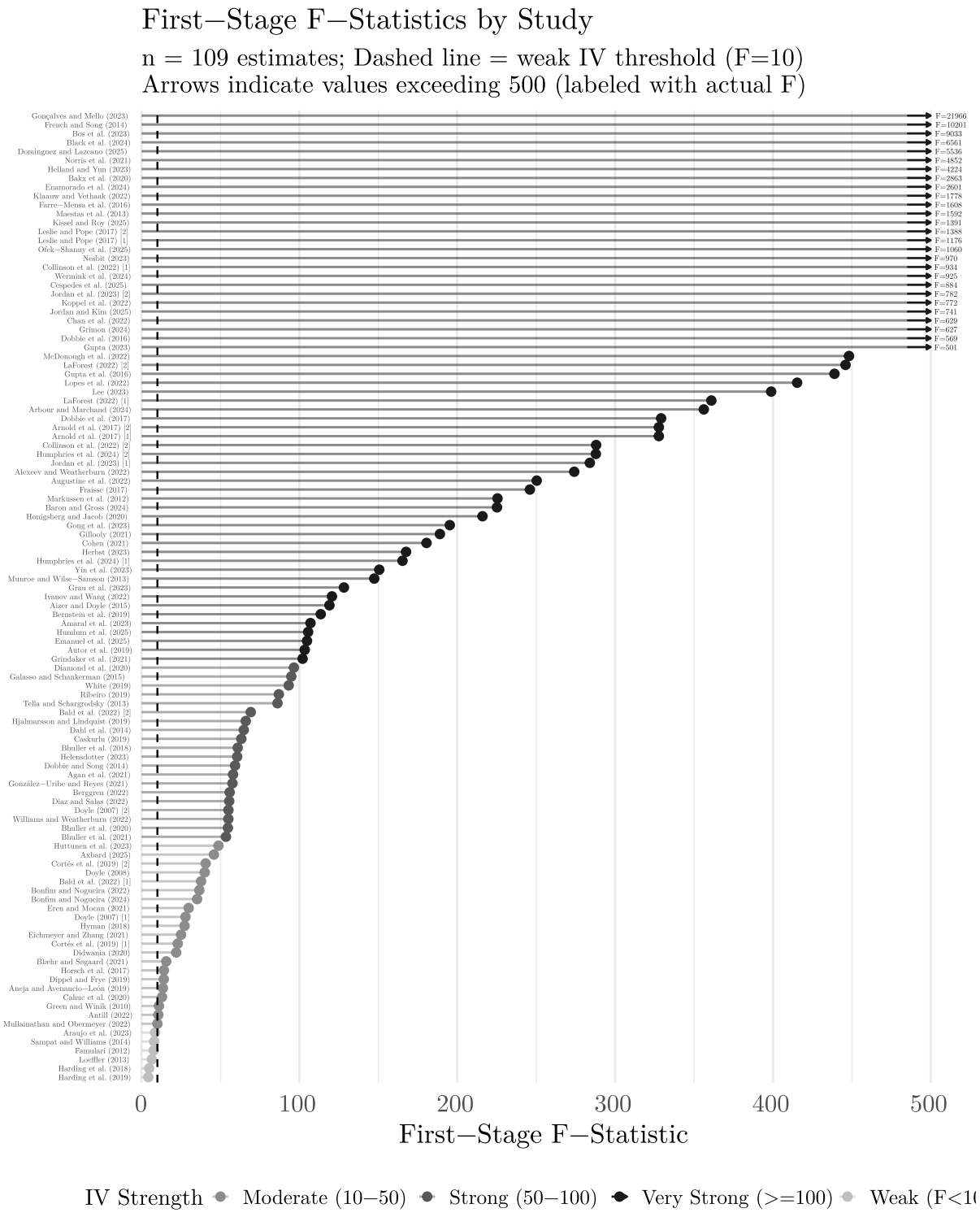


Figure 38: First-stage F-statistics by study, ordered by magnitude. Dashed line marks the weak-instrument threshold ($F = 10$). Arrows indicate values exceeding 500, labeled with the actual F-statistic.

D.9 How selected is the leniency-IV sample?

Ideally, I would like to generalize from the settings in the leniency-IV sample to a broader range of bureaucratic decisions. However, the studies are selected. The sample can therefore misrepresent bureaucratic decision-making through three channels. (1) The design only works where examiners disagree enough to clear the weak-instrument convention of $F \geq 10$, so settings with little disagreement may be missing from the literature. (2) Authors choose which subsample carries the headline estimate. And (3) researchers choose which settings to study at all, plausibly drawn toward disparities that were already public.

How much dispersion does a study need before it can exist? Since they are instrumental variable designs, we can think of each study as having a threshold, the smallest true dispersion that would still give an expected first-stage F of 10, which is the conventional saying that this is the minimum, given the study's caseload per examiner, number of examiners, base rate, and the type of F it reports. Figure 39 plots noise-corrected GMDs against caseload with the MDE thresholds, and Figure 40 shows the ratio of each study's corrected dispersion to its own threshold. The median study has 3.6 times as much GMD as it could have had and still be a powerful enough instrument (IQR: 2.4 to 6.9), and the median threshold is a GMD of 1.6 percent, so at the typical scale of the studies, the requirement can only exclude settings where a different draw changes very few decisions. It is not strictly enforced either, which helps the findings generalize beyond the sample. In 9 of 100 studies, the number is below their own threshold, and there is no bunching around a reported F of 10 (three main estimates in $[7.5, 10)$ against three in $[10, 12.5)$; the same counts hold across all 201 estimates).

What if low-dispersion settings were never written up? Because the weak-instrument problem mimics publication-bias logic, in which too little power prevents a study, I can test sensitivity in the same way. The selection model treats log dispersion across settings as normally distributed and lets a setting below its own threshold enter the literature with probability π_0, estimated from the data. This is the publication-bias selection model of Andrews and Kasy (2019), with two changes fitted to this application: selection acts on the design's expected first-stage F rather than on the significance of a headline estimate, and the threshold is known from the weak-instrument convention rather than estimated. With the estimated selection ($\hat{\pi}_0 = 0.19$: a below-threshold setting is about one fifth as likely to be written up), the population median GMD is 5.2 percent against 5.8 percent in the observed sample, with a bootstrap 95 percent interval of 4.2 to 6.4. Assuming instead that below-threshold settings are never written up at all, and refitting on the studies above their thresholds, gives 4.8 percent; raising the threshold to $F = 20$ gives 5.1 percent. The convention of 10 plays the same role here as the 5 percent significance bar in the original model, and refitting with the bar at 15, 20, 30, 50, and 104.7, the smallest first-stage F at which the usual 5 percent confidence interval is valid without adjustment for a single instrument (Lee et al., 2022), moves the corrected median from 5.2 down to about 4.8 at a bar of 30, where it stays through 50 and 104.7 (Figure 41). At a bar of 104.7, one published study in five sits below it, so the model concludes such a bar is not very

applicable ($\hat{\pi}_0 = 0.25$).

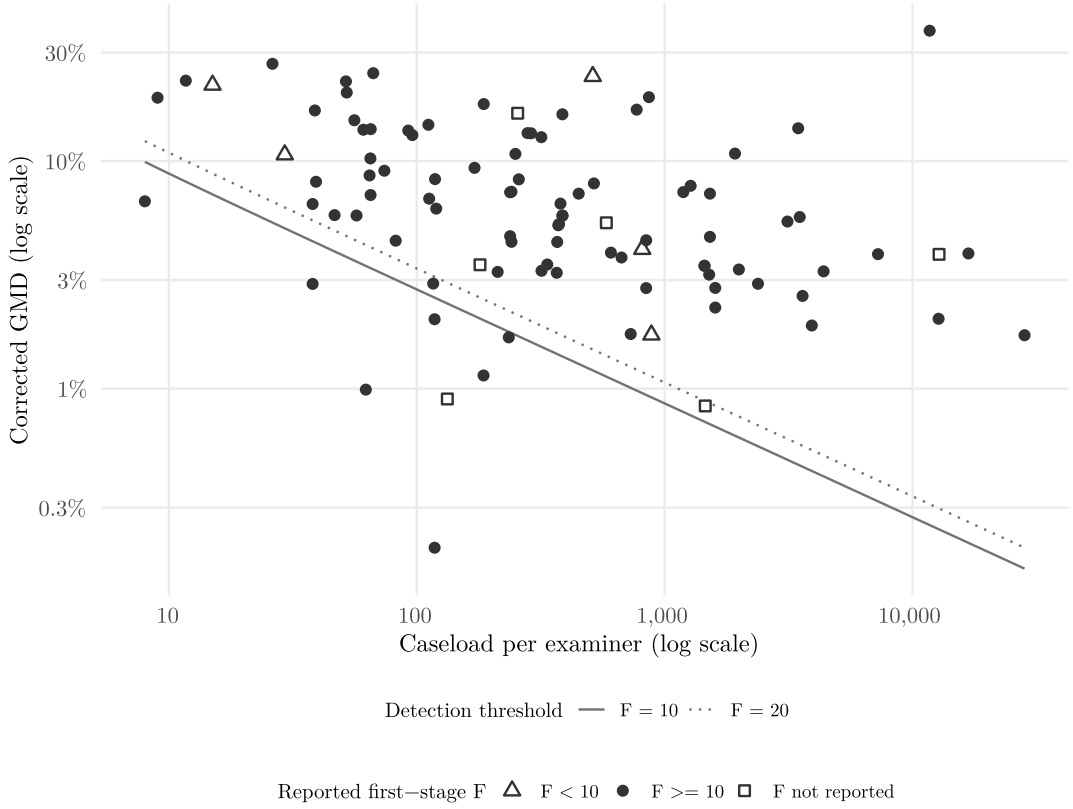


Figure 39: Corrected GMD against caseload per examiner, with detection thresholds. The vertical axis is the noise-corrected GMD of Equation 21, the paper's primary estimand, not a selection adjustment. Curves show the smallest true GMD that gives an expected first-stage F of 10 (and 20) at the sample-median base rate and examiner count, for the single leave-out instrument that all but four studies use; a study above the solid curve could have existed under the convention, and nearly every study sits well above it. The four studies reporting joint examiner-dummy Fs face a higher bar, which is used in the calculations. Triangles mark studies whose reported F is below 10; squares mark studies with no reported F. Three studies with a corrected GMD of exactly zero do not appear (two outside the log scale, one without usable caseload data); they enter all counts.

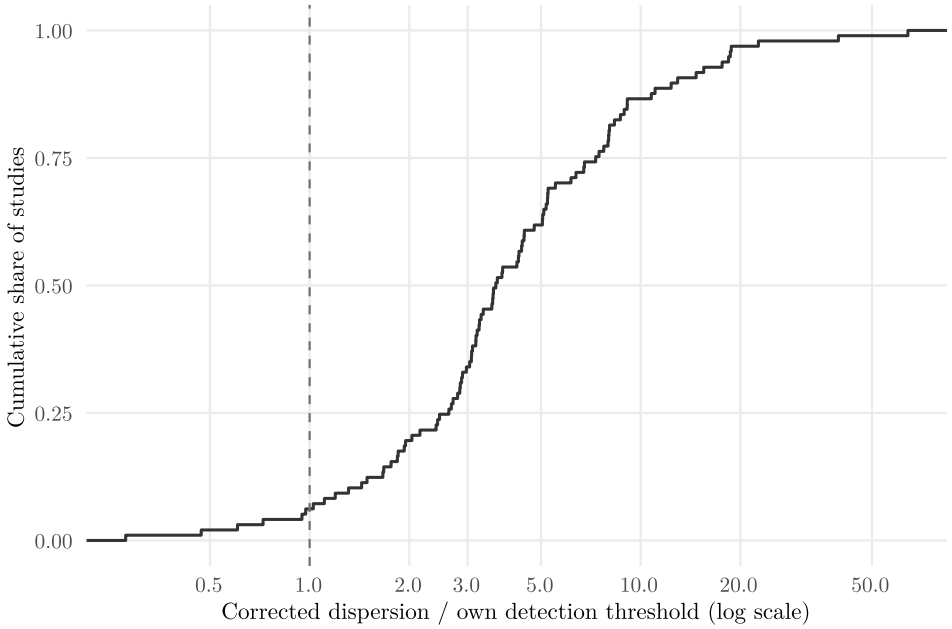


Figure 40: Cumulative distribution of each study's corrected dispersion divided by its own detection threshold at $F = 10$. The dashed line marks a ratio of 1, the point where the F requirement would begin to bind. Three studies with corrected dispersion of exactly zero are outside the log scale and omitted.

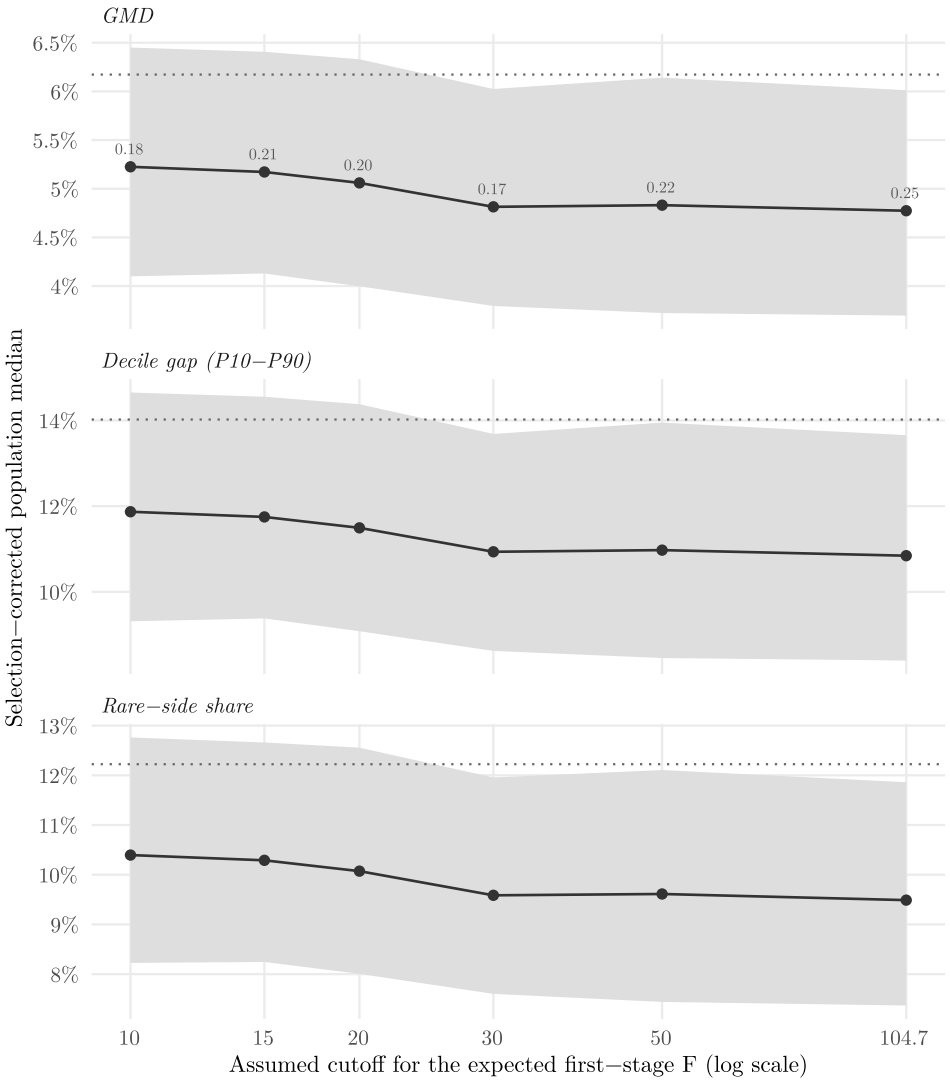


Figure 41: Selection-corrected population medians of the three estimands as the assumed publication bar rises. Bands are bootstrap 95 percent intervals.

Is the caseload gradient a selection artifact? Small-caseload systems show more dispersion than large-caseload systems (Spearman −0.37; subsection D.18), and a selection story could explain the same pattern. The detection threshold sits high where caseloads are small and low where they are large, so if settings below their threshold never become studies, the lower-left corner of Figure 39 is empty by construction, and the surviving studies slope downward even if true dispersion had nothing to do with caseload. Two checks tell these stories apart. First, if the pattern came from the threshold, it should live in the studies close to their thresholds and weaken when they are set aside. It strengthens instead: the relation is −0.39 among studies with $F \geq 50$ ($n = 69$), −0.46 at $F \geq 100$ ($n = 52$), and −0.52 among studies with at least twice the dispersion their own threshold required ($n = 78$). Second, the noise correction could draw the pattern. If the correction removed too little noise, the leftover would be largest in small-caseload studies, because sampling noise grows as caseloads shrink, and that alone would tilt the picture downward. A simulation checks this: give every study a true dispersion unrelated to its caseload, redraw its data at its own size, and run it through the same correction as the real pipeline. The

simulated relation centers on +0.01 (95 percent band −0.19 to +0.22), none of 500 runs reaches the observed −0.37, and the correction recovers the truth within 9 percent in each fifth of the sample sorted by caseload. The pattern is a property of the settings, not of the selection and not of the correction.

Do authors pick their most dramatic context? If authors pick the context with the most dramatic dispersion as the headline, headline contexts should have higher arbitrariness than secondary contexts reported in the same papers. On corrected GMDs they do not: the main context is higher in 8 of 15 papers with usable pairs, a coin flip (sign test $p = 1.00$). Raw, uncorrected measures point the opposite way (main higher in 2 of 16 papers on the P10–P90 range, $p = 0.004$, and 3 of 18 on the raw SD, $p = 0.008$), because secondary contexts are smaller subsamples whose raw spread is inflated by sampling noise; the reversal after correction is a noise artifact, not evidence of reporting selection.

Which settings can never appear in the sample? I can address the remaining channel, which settings get studied at all, only with a scope statement. I cannot report on two types of excluded settings. Settings with a single bureaucrat have no other bureaucrat to compare with, so the estimand is not defined there. Settings without quasi-random assignment are outside the design's reach in either direction, and where assignment is steered on purpose, a reassignment is no longer luck, so the estimand would mean something else.

Did researchers pick the settings with the most disagreement first? The worry is this: researchers may choose to study a setting because its bureaucrats are already known to disagree. If that is how settings get chosen, the published sample holds the settings with the most disagreement, the settings nobody studied hold less, and my headline overstates the ordinary setting. I cannot measure the unstudied settings directly. But choosing happens one study at a time, so it would leave a track in the order the studied settings arrived. If researchers had always gone for the setting with the most disagreement first, the early literature would hold the biggest cases, and each year's new settings would show less disagreement than the year before, because the biggest ones get used up. I check for that. I sort the 118 estimates into 83 systems (one institution making one kind of decision) and give each system the year of its first study. Figure 42 plots each system's corrected GMD from that first study against that year. The line is flat: systems that entered in 2024 and 2025 show the same disagreement as systems that entered between 2007 and 2013 (Spearman −0.03, $p = 0.80$; least-squares slope −0.11 percentage points per year, standard error 0.20; 79 systems). The new arrivals are not leftovers. The flat line is also not about what studies could measure, because every year's newcomers found far more disagreement than the smallest amount their studies could have detected. And counting each system once leaves the median at 5.9 percent against 5.8, so the systems studied many times are not special. One picking rule escapes this check: if researchers only ever study settings above some level of disagreement, in any year, the order of arrival shows nothing. The full populations below test that rule directly.

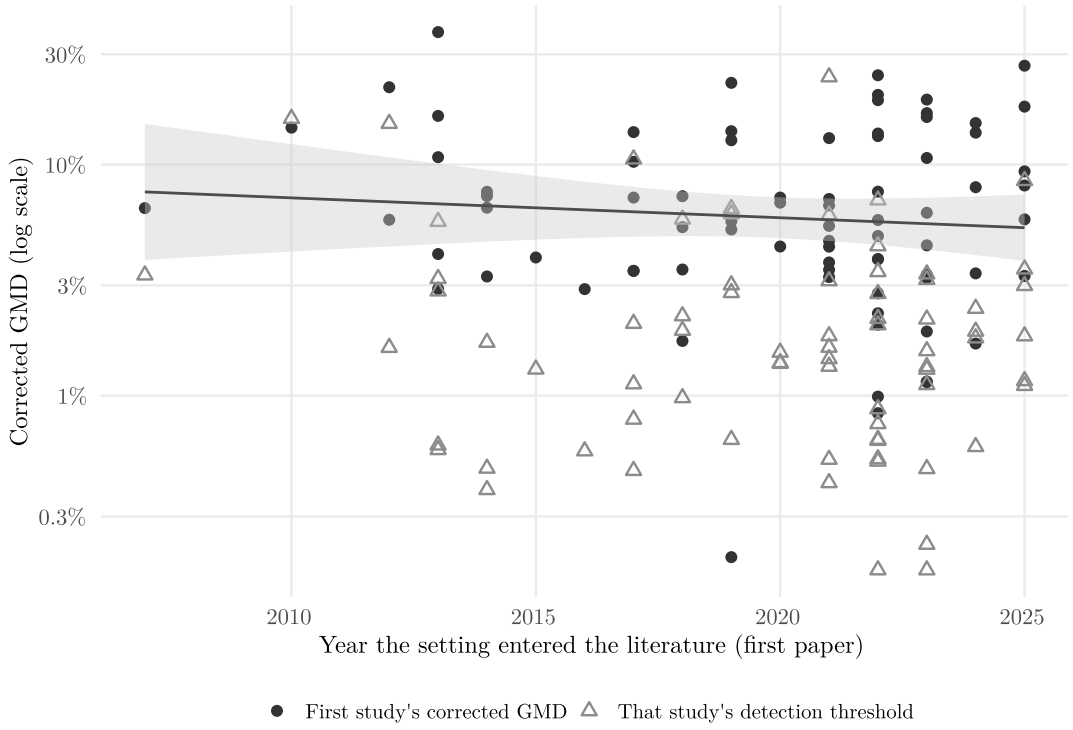


Figure 42: Corrected GMD of each decision-making system's first study, by the year the system entered the literature (filled circles), with that study's detection threshold (open triangles). The line is the least-squares fit with its 95 percent band. Recent entrants show the same dispersion as early entrants, and entrants sit well above their thresholds throughout. Two systems whose first study has a corrected GMD of zero are outside the log scale and omitted.

What would we find where nobody selected the setting? The strongest evidence on site choice comes from full populations where nobody selected on the instrument. I computed the paper's corrected GMD in nine complete populations of bureaucrats, assembled by data availability and processed under a protocol that fixed the benchmark set, margins, and estimator before any comparison was computed (Table 22). Four of the systems already had leniency-IV papers; five have no dispersion literature at all. The Kenya data come from Choi et al. (2022), and the Israel data come from Shayo and Zussman (2011). The censuses of the nine settings do not sit below the published studies' summary statistics: corrected GMDs run from 6.7 to 29.0 percent, and the residualized versions, which compare bureaucrats only inside the same court, art unit, district, origin country, or county and year and are the closer analogue to a published leniency instrument, run from zero in Kenya, reflecting thin data within cells rather than agreement, to 24.1 percent in Austria, against the paper sample's median of 5.8 percent.

Where I could get data on settings that have already been studied, the numbers match or exceed what the papers extracted: patent examiners 17.0 percent against a published patent-domain median of 15.8; federal sentencing 5.9 against 5.2; disability allowance 14.3 (an aggregate release, no residualization possible) against 7.3. The domain with the largest dispersion, asylum adjudication in three countries, hosts no paper in the meta-analysis sample at all, but is known in the literature to have very high numbers (Ramji-Nogales et al., 2007; Rehaag, 2008; Byrne et al., 2024). If site choice were pulling the headline numbers upward, these unselected populations

should sit below the published sample, but on the corrected measures they sit at or above it everywhere, so the published literature is actually just a middle draw from the feasible number of settings worldwide, not its upper tail. There are some estimation caveats, but this is mostly meant to show that selection is not very apparent.

Table 22: Census benchmarks: dispersion in full populations of bureaucrats.

System	Margin	Bureaucrats	Cases	Base	GMD	GMD^{res}	Published median
Systems with published leniency-IV papers							
US EOIR asylum	asylum grant	1152	912,911	40.0	24.5	14.1	–
US SSA ALJs	allowance	3552	8,721,978	45.8	14.3	–	7.3 (n=5)
US federal sentencing	incarceration	1046	547,353	89.2	9.3	6.0	5.2 (n=14)
USPTO patents	allowance	11089	2,656,373	63.1	24.5	17.0	15.8 (n=5)
Leniency-ready systems without a dispersion literature							
Austria asylum court	protection grant	350	82,233	50.1	29.0	24.1	–
Canada RPD asylum	merits grant	579	148,614	55.9	25.2	17.5	–
Israel small-claims	claim accepted	58	1,596	73.9	8.8	4.9	–
Kenya High Court appeals	pro-defendant outcome	139	10,465	56.2	13.4	0.0	–
Slovakia courts	penal order	428	208,818	67.4	18.3	16.3	–
Slovakia courts	appellate disposal	1110	607,654	5.0	6.7	5.6	–

Notes: Each row is the full population of bureaucrats in one system, assembled by data availability, with no first-stage or publication filter. GMD is calculated as I described in the paper. GMD^{res} residualizes decisions within cells before the same correction (EOIR: court × year; USPTO: art unit × year; sentencing: district × year; Canada: origin country × year; Austria: year; Slovakia: court × year; Israel: court; Kenya: county × year). SSA is an administrative per-judge aggregate release, so no residualization is possible. Kenya's residualized value is set to zero, because county-year cells are small (16 percent of cases sit in single-judge cells), so the within-cell signal sits below the noise level; it does not demonstrate agreement. In Austria the "bureaucrats" are court divisions. Slovakia margins are constructed from order types.

Is Study 2 selected the same way? Conjoint papers are selected on average attribute effects, not on how much respondents disagree; the split between shared and idiosyncratic weights is a by-product that no editor screens on.

D.10 Additional conjoint results

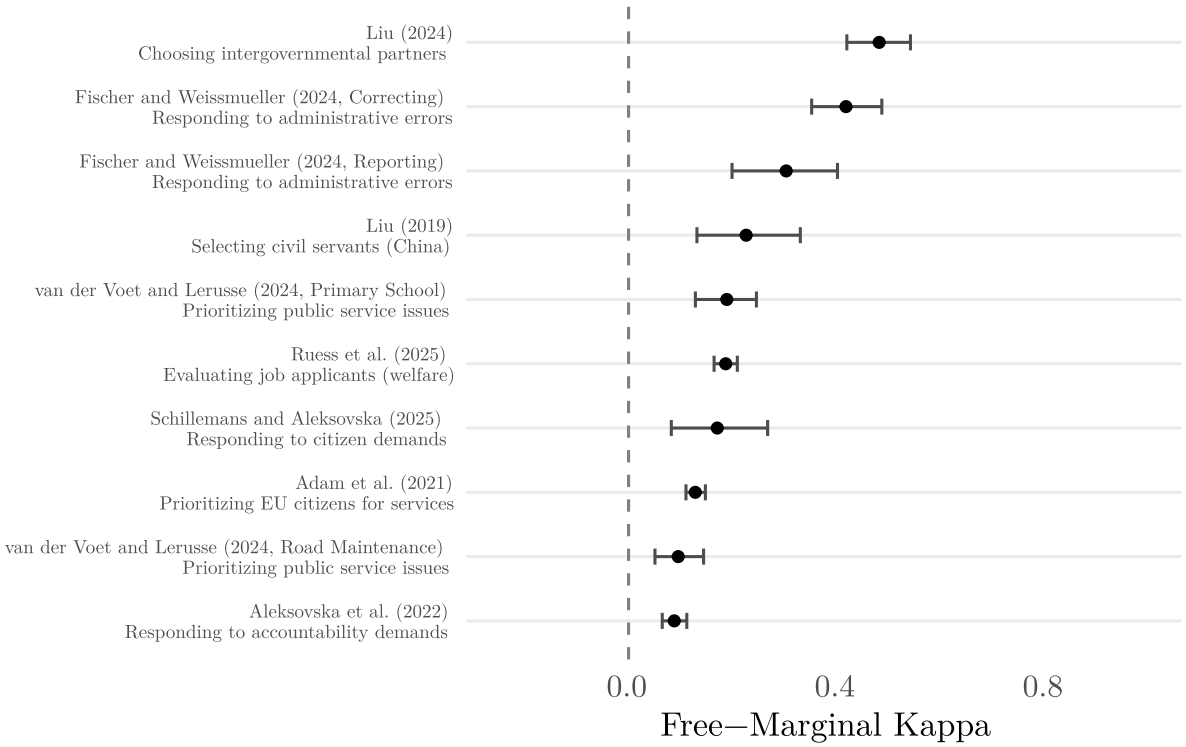


Figure 43: Free-marginal kappa across conjoint studies. Kappa is weighted by the number of pairwise comparisons per shared vignette. Error bars are 95% bootstrap intervals from 2,000 comparison-level resamples. Values near zero indicate agreement no better than chance; values near one indicate perfect agreement.

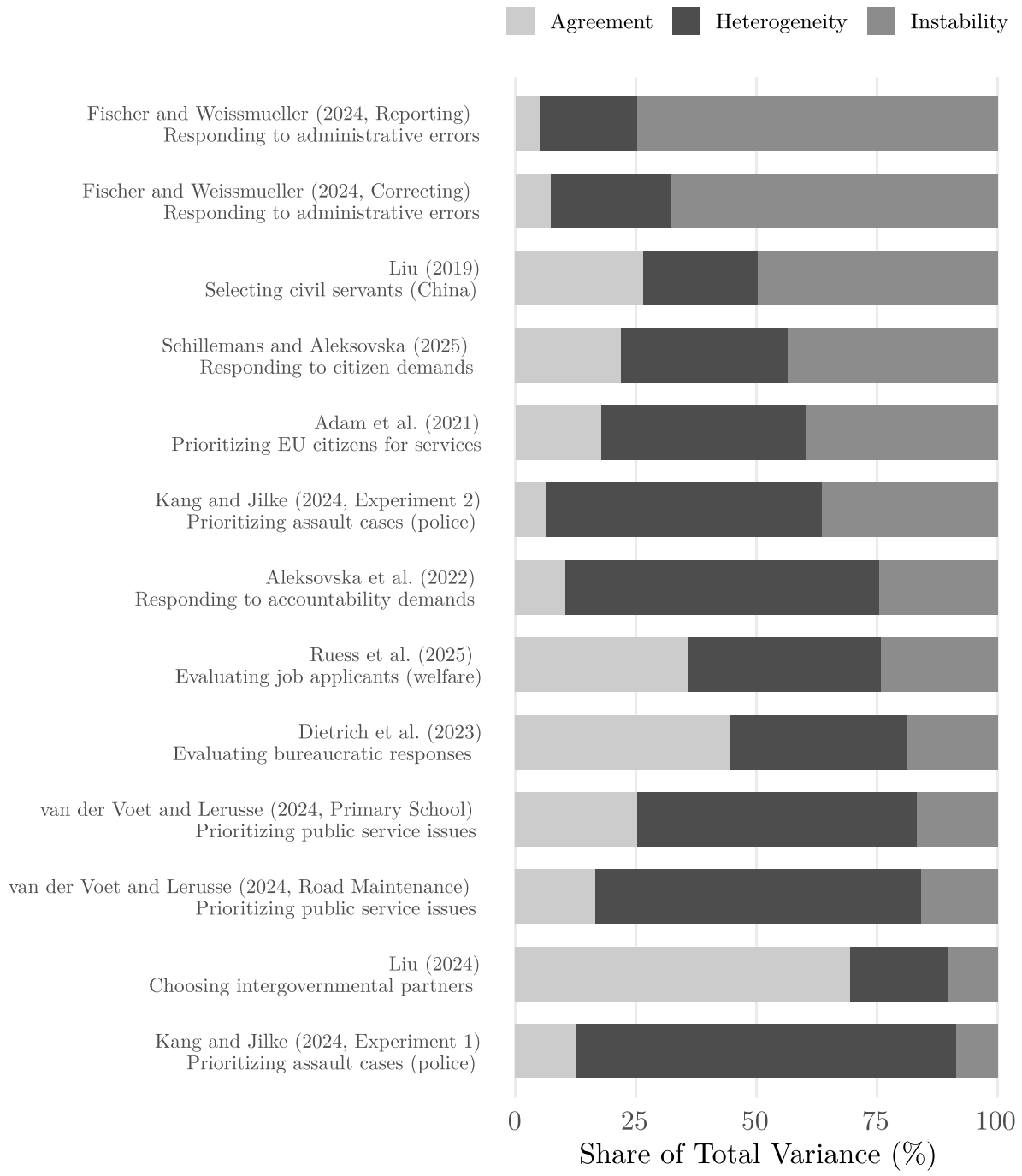


Figure 44: Full variance decomposition of conjoint responses, ordered by instability share. Agreement captures shared attribute preferences; heterogeneity captures stable between-respondent differences; instability captures within-respondent noise.

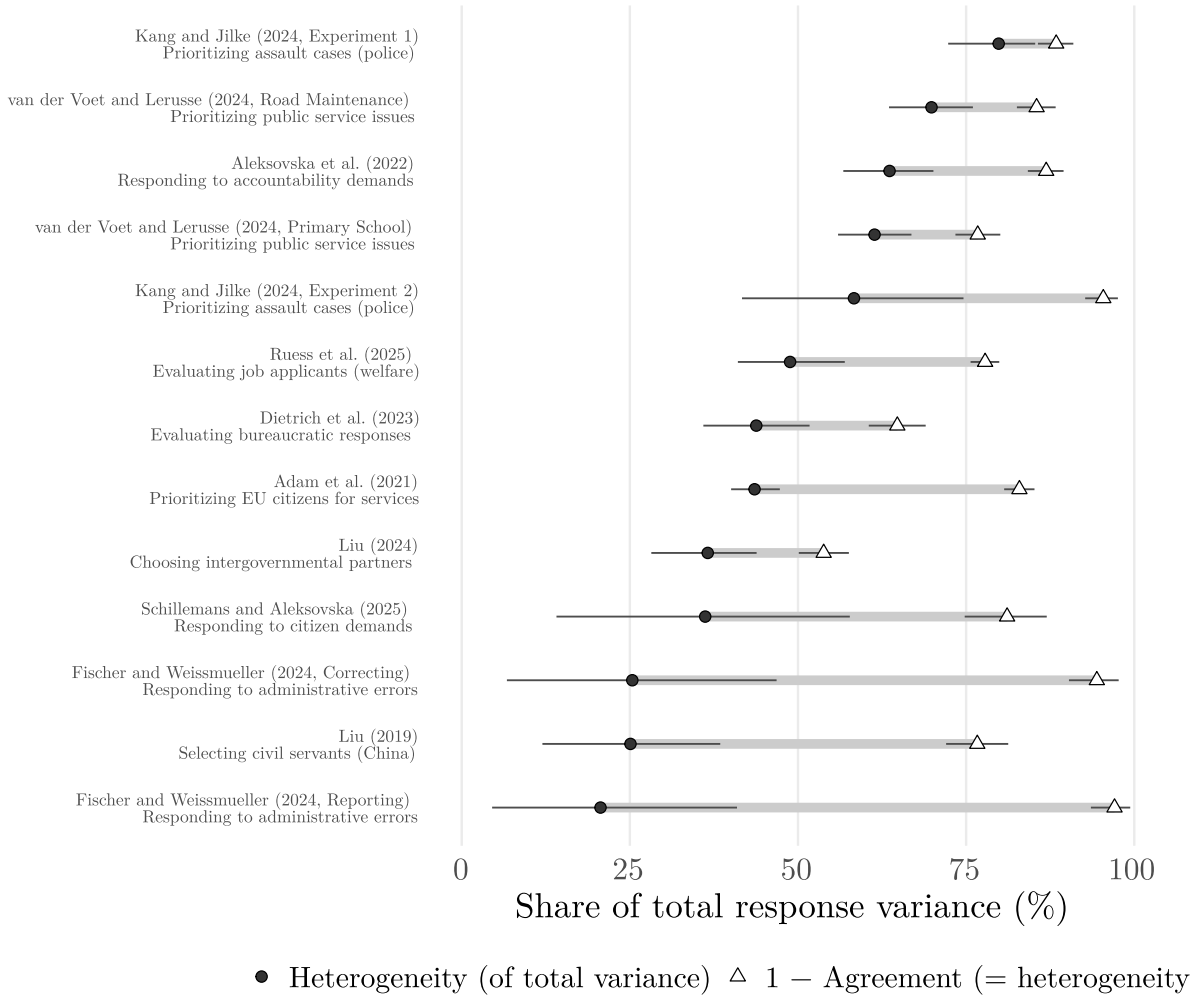


Figure 45: Variance decomposition on the total-variance scale, by estimate, with 95% posterior credible intervals. The filled circle is heterogeneity as a share of total response variance; the open triangle is one minus agreement, all variance that is not shared consensus (heterogeneity plus instability); the grey span between them is the instability (within-respondent noise) share. Each interval is drawn on the line of its own point. Estimates are ordered by heterogeneity. The main-text Figure 16 shows heterogeneity renormalized to the systematic-variance scale.

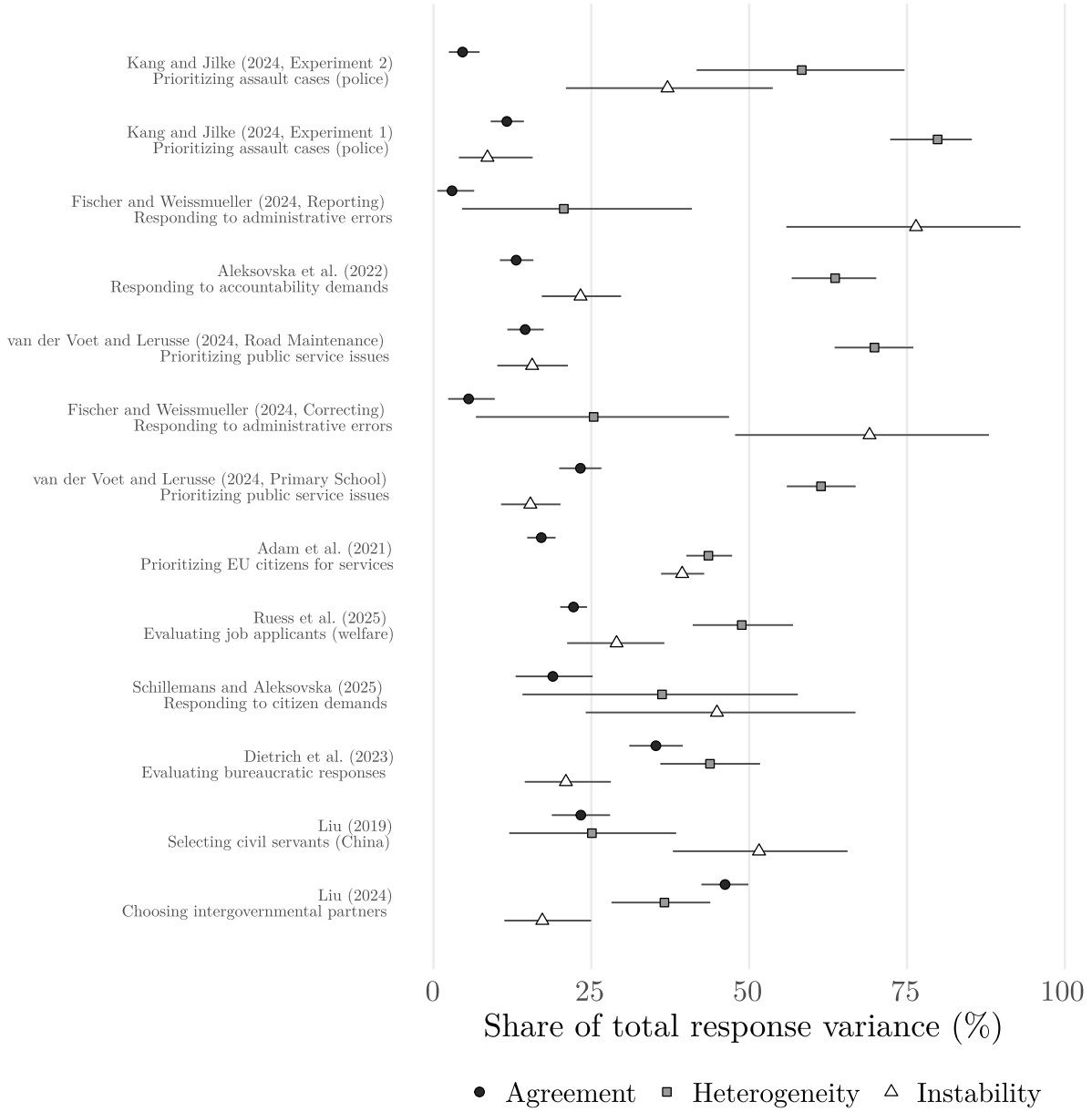


Figure 46: All three variance shares in one panel (companion to Figure 17). Points are posterior-mean shares of total response variance for each estimate; horizontal bars are 95% credible intervals. The black circle is agreement, the grey square heterogeneity, and the open triangle instability. Estimates are ordered as in Figure 16.

D.11 Subgroup comparisons

Table 23: Arbitrariness by interaction mode (non-judge bureaucrats only).

	Range (P10–P90)		β^{std}		Relative arbitrariness	
	Median	Mean	Median	Mean	Median	Mean
Face-to-face ($n = 34$)	17.6%	20.0%	3.2pp	5.4pp	16.5%	28.3%
Remote ($n = 14$)	23.1%	28.2%	6.9pp	6.5pp	17.8%	23.6%

Wilcoxon rank-sum: range $p = 0.091$; β^{std} $p = 0.190$.

Table 24: Arbitrariness by decision type. Sanctioning includes incarceration, detention, arrest, eviction, fines, and sanctions. Supporting includes benefit approval, treatment, housing assistance, and debt relief. Administrative decisions include patents, diagnostics, child welfare assessments, and procedural determinations.

	n	Range (P10–P90)		β^{std}		Relative arbitrariness	
		Median	Mean	Median	Mean	Median	Mean
Sanctioning	56	13.1%	20.7%	3.4pp	5.0pp	12.0%	18.3%
Supporting	36	17.4%	22.3%	5.7pp	5.6pp	18.6%	26.8%
Administrating	26	19.2%	23.9%	3.3pp	6.2pp	20.8%	29.1%

Table 25: Arbitrariness by study characteristics.

	n	Range (P10–P90)		β^{std}		Relative arbitrariness		Wilcoxon p
		Median	Mean	Median	Mean	Median	Mean	(range/β)
Bureaucrat type								
Judge	69	14.5%	21.2%	3.9pp	5.3pp	15.8%	21.0%	0.334 / 0.901
Non-judge	48	19.2%	22.3%	4.3pp	5.7pp	17.5%	27.1%	
Topic								
Criminal justice	54	12.9%	18.0%	3.3pp	4.4pp	12.6%	20.8%	0.030 / 0.049
Other	62	21.0%	25.5%	5.6pp	6.3pp	18.6%	25.9%	
Country								
US	79	14.1%	21.0%	3.5pp	5.4pp	15.4%	22.5%	0.190 / 0.327
Non-US	38	17.6%	23.0%	4.9pp	5.5pp	20.8%	26.0%	
Legal system								
Common law	86	14.8%	21.2%	3.8pp	5.5pp	16.2%	22.4%	0.289 / 0.560
Civil law	31	17.0%	22.8%	4.8pp	5.4pp	20.8%	27.1%	

D.12 Estimation details

There are some practical difficulties in calculating the actual leniency part of the 2SLS the studies report on. First, sometimes in the studies they choose to focus on a smaller set of decisions in the regression tables, but they calculate the leniency, which is the number I want from a larger set, before the actual 2SLS estimation. So, the caseload $\bar{n}$ must come from the sample the study used to build its leniency measure, which is often much larger than its analysis sample, and if I used the wrong one, I remove too much, or I would ascribe too much to sampling noise. Second, nine studies report no usable caseload. There I read the reliability off the study's own first stage: noise pulls the first-stage β of the design equation below one, and β estimates the share of the measure's variance that is signal, so $\hat{\sigma} = \text{SD}_{\text{obs}}\sqrt{\hat{\beta}}$. Third, when caseloads are very unequal the average caseload slightly understates the noise, which moves the spread by at most five percent (see Appendix D.13). Fourth, studies differ in the minimum caseload below which they drop bureaucrats before reporting a spread; I record each threshold, and the correction puts studies with different thresholds on the same noise-free scale.

Also, there is a difference between raw and residualized spreads. Residualized results are biased slightly lower than the raw ones, because residualizing removes real differences along with the controls. I do not have to subtract the sampling noise when the studies already use shrunken, fitted, and grouped spreads. Where assignment is random only within strata, the reported spreads are computed within those strata as well; that is the variation the lottery actually runs on, so it is the right comparison.

The whiskers on the study-level figures are parametric bootstrap intervals, one simulation per study on the moment route. I do not hold the case files, only four numbers per study: the number of bureaucrats J, their average caseload $\bar{n}$, the base rate p, and the corrected spread $\hat{\sigma}$. Each of 2,000 replications redraws the study from those numbers: draw J true decision rates from a normal distribution with mean p and spread $\hat{\sigma}$, give each bureaucrat $\bar{n}$ cases decided at their true rate, take the spread of the resulting observed rates, and re-run the correction of equation (21) on it. The middle 95 percent of the corrected spreads, re-centered on the study's point estimate through the simulation median, is the interval for σ. Because the estimands are fixed multiples of σ, the same interval scales into all of them: $1.128\,\sigma$ for the GMD, $2.563\,\sigma$ for the decile contrast, and GMD over $2p$, capped at one, for the rare-side share; the base rate is treated as known. The interval therefore carries both layers of sampling noise in the estimator: which bureaucrats the study happened to contain and the caseload noise in each estimated rate. Slope-route and lower-bound rows carry no interval, because this simulation is the moment estimator's own and would misstate their uncertainty; direct-route rows inherit the converted interval rescaled by the ratio of the direct to the converted estimate, an approximation.

The interval on the cross-study median is a different bootstrap, nonparametric this time. I resample the 88 papers with replacement, keep every estimate of each resampled paper, recompute the median, and repeat 2,000 times; the middle 95 percent of these medians is the reported interval. Resampling whole papers rather than single estimates respects that estimates from the same paper share data and methods. Both procedures run from fixed seeds, so the intervals reproduce exactly.

D.13 Getting the numbers without replication data

Most studies in the meta-analysis publish no replication data, so a fair question is how I can compute a corrected GMD and a complier share for them at all. The answer has two parts. First, every number I use is taken from the study itself: its printed text and tables, its published figures, or, for two studies, numbers the authors sent on request. Any value in any figure of this paper can be traced to its source, because Table 31 lists each study's inputs, value-by-value, with the page each one came from. Second, the routes that turn published material into my numbers can be tested, because for some studies the truth is known: six studies publish their raw case files, 21 print exact percentiles of their leniency distribution, and one author team sent the full distribution itself. This appendix reports those tests.

Each route is graded against every kind of truth it can meet (Table 26). Reading a figure bar by bar recovers exactly stated percentiles to within about 1 percent of the study's spread, and stated percentile gaps, the observed complier share, to within about 2 percent. Converting a standard deviation through the normal distribution lands within about 10 percent of stated gaps and leans high, because real leniency distributions bunch more than a normal distribution; for the 14 studies where I hold a full distribution I therefore compute from the distribution itself and use the normal shape only as the fallback. The strongest test rebuilds everything from published summary numbers alone for the six case-file studies and runs the same computation

on the raw cases: the answers differ by 5 percent at the median and 16 percent at worst, and the worst gap is 0.004 in absolute terms, on the smallest spread in the set (Table 27). The correction step itself adds none of this error: fed the exact case-file inputs it reproduces the case-file answer to within half a percent, and a split of each bureaucrat's cases in two, which uses no formula, recovers the same spread. The one author-sent distribution brackets my converted GMD within two percent (0.161 against a band of 0.157 to 0.163).

The comparisons compare like with like. Where a study's own construction defines a different object, such as a rate measured within a court, I compute that object from the cases, and the note column of Table 27 records what each remaining gap traced to. Papers that publish case files or exact percentiles may be unusually careful, and coverage is not selective on dispersion: the median observed spread is 0.17 among all estimates and 0.16 among those with corrected numbers.

Table 26: The recovery tests. Each row feeds one route only what a paper publishes and compares the result with a known truth. Gaps are percentage points of the study's spread.

Route	Truth it is tested against	Checked on	Result
Figure, read bar by bar	the same percentile stated exactly in the paper	20 pairs, 9 figures	median 1%; worst 21%, an extreme tail percentile
Figure, read bar by bar	the stated percentile gap (the observed complier share)	6 gaps, 5 studies	median 2%; worst 28%, which is 0.009 on a near-zero gap
SD converted through the normal distribution	the stated percentile gap	8 gaps, 7 studies	median 10%, leaning high; worst 46%, a study whose inputs are under object review
Everything at once, from published summary numbers	the full computation repeated on the raw case files	6 studies	median 5%; worst 16%, which is 0.004 in absolute terms (Table 27)

Table 27: The end-to-end test compares, for the six studies with public case files, the corrected spread built from published summary numbers alone with the same computation run on the raw cases. Gaps are percentage points of the case-file spread.

Study	From the paper	From the cases	Gap	Note
Green and Winik (2010)	0.133	0.126	6%	printed SD matches raw files to 0.3%
White (2019)	0.032	0.030	6%	matches; caseload convention corrected
Di Tella and Schargrodsky (2013)	0.031	0.030	3%	two input mislabels found, corrected
Agan et al. (2023)	0.029	0.025	16%	within-court object, matched via the paper's own residualization
Humphries et al. (2025)	0.030	0.029	5%	windowed leave-out object, matched via the paper's own construction
Axbard (2025)	0.217	0.227	5%	instrument SD had been filed as the distribution SD; corrected (figure percentiles match the files within 3%)

D.14 Where each number comes from

Table 28: Which numbers I have, where they come from, and how they are checked.

Input or check	Estimates	Share of 118 or result
Where the spread of the bureaucrats' decision rates comes from		
the study prints a standard deviation, or I read one off its figure	101	86%
the study prints percentiles; I turn the percentile gap into a standard deviation	7	6%
I compute it from other numbers the study prints	2	2%
no spread is reported anywhere; the study is excluded, its reason listed	8	7%
The two numbers that set how much spread sampling noise alone would create		
share of cases granted: printed in the study	109	92%
share of cases granted: taken as the average of the bureaucrats' rates	4	3%
share of cases granted: read off a figure or taken from replication files	4	3%
share of cases granted: missing (the study is among the excluded above)	1	1%
cases per bureaucrat: counted in the sample the rates were built on	99	84%
cases per bureaucrat: only the analysis sample's count is available	11	9%
How the noise is removed, study by study		
subtracted with the formula, from the share granted and the cases per bureaucrat	84	71%
taken from the study's own regression of decisions on the leniency measure	8	7%
nothing subtracted: the study already removed the noise, so its spread is a lower bound	4	3%
no formula: everything is computed from the full distribution of rates	14	12%
How the numbers are checked (one study can appear in several rows)		
six studies publish their raw case files		redoing every step from the cases lands within 5% of my answer (median)
9 figures re-read bar by bar, where the paper also prints exact values		the re-reading matches the printed values within 1% (median)
the original figure readings (14 figures can be checked)		within 9% of the spread (median); replaced wherever a better source exists
21 studies print exact percentiles		these printed values are the measuring stick for the two figure checks above

D.15 Raw vs. residualized instruments

Table 29: Raw vs. residualized leniency instruments: comparison of arbitrariness measures

Measure	Raw instrument		Residualized instrument		t-test
	n	Mean	n	Mean	p-value
Range p_{10}–p_{90}	57	0.229	43	0.219	0.773
β^{std}	59	0.054	41	0.046	0.308

Notes: Studies are classified as "residualized" if the reported mean of the leniency instrument is within 0.01 of zero, indicating that necessary controls have been partialled out; otherwise, they are classified as "raw." p-values are from two-sided Welch t-tests.

D.16 Web of Science Search Query for Leniency-IV Studies

The following Topic Search (TS) query was used on Web of Science to identify studies using the leniency instrumental variable design:

```
TS=("examiner leniency" OR "judge leniency" OR "caseworker leniency"
OR "inspector leniency" OR "physician leniency" OR "teacher leniency"
OR "examiner stringency" OR "judge stringency"
OR "examiner fixed effect*" OR "judge fixed effect*"
OR "randomly assigned judge*" OR "randomly assigned examiner*"
OR "randomly assigned caseworker*"
OR "judge instrumental variable*" OR "examiner instrumental variable*"
OR "as-good-as-random assignment"
OR ("leave-out mean" AND ("judge*" OR "examiner*" OR "caseworker*"
    OR "inspector*" OR "physician*"))
OR ("jackknife" AND ("judge*" OR "examiner*" OR "caseworker*"
    OR "inspector*"))
OR ("quasi-random assignment" AND ("judge*" OR "examiner*"
    OR "caseworker*" OR "inspector*" OR "physician*"))
OR ("leniency" AND "instrumental variable*")
OR ("stringency" AND "instrumental variable*")
OR ("tendency" AND "instrumental variable*" AND ("judge*"
    OR "examiner*" OR "caseworker*")))
```

Conjoint studies from Web of Science search:

D.17 Re-studied systems

Several systems in the pool have been studied more than once, and I count each system once. When two papers study the same bureaucrats making the same decision, I keep the newest

paper, or the older one if its data quality and reporting are significantly better. At the screening stage this removed four papers: Doyle (2013), which reuses the Illinois investigator design of Doyle (2007) and Doyle Jr. (2008); Gross and Baron (2022), which uses the same Michigan investigators as Baron and Gross (2025); Lee (2022), an earlier version of Lee (2023); and the working-paper version of Bos et al. (2023). Within the extracted sample the rule removes 15 estimates from 13 papers. Keeping the newest: Bonfim and Nogueira (2024) over Bonfim and Nogueira (2022); Bhuller, Khoury and Løken (2021) over Bhuller et al. (2018) and Bhuller et al. (2020); Autor et al. (2019) over Dahl et al. (2014); Black et al. (2024) over French and Song (2014); Harding et al. (2019) over Harding et al. (2018); Dobbie et al. (2017) over Dobbie and Song (2015); Jordan et al. (2025) over Loeffler (2013), which studies the same Cook County assignment computer over 2000–2003, inside the newer paper's 1990–2007 window; Enamorado et al. (2024) and McDonough et al. (2022) over the two city estimates of Arnold et al. (2018), whose reported dispersion pools the Miami-Dade and Philadelphia lotteries; Koppel, Bergin, Ropac, Randolph and Joseph (2022) over Leslie and Pope (2017), whose misdemeanor assignment fails its own randomization test, and over Kleinberg et al. (2018), which reports only grouped leniency-quintile means on the same 2008–2013 case file that Leslie and Pope use; and Dominguez and Lazcano (2025) over Grau et al. (2023), the same national Chilean judge lottery measured on the per-assigned-case rather than the per-request margin. Keeping the older paper under the quality clause: Horsch et al. (2021) over Caskurlu (2019). The two tap the same randomly composed appellate panels on different case universes and margins, and only the older paper reports the judge-level distribution, a named per-judge table with propensities and caseloads, while the newer reports a single instrument dispersion that enters my data as a lower bound. For every removed pair with a different author team I verified against the papers that the two study the same assignment lottery: same institution, same assignment mechanism, same decision margin, and overlapping case universes; the per-pair audit sits in the replication files. Dropping these estimates changes the three medians by at most one percentage point.

D.18 Caseload correlation: mechanical relationship check

If the number of examiners in a study is large, two things happen simultaneously: (a) average caseload per examiner falls (holding total cases fixed), and (b) the estimated distribution of examiner propensities is better sampled, potentially producing wider observed ranges or more precise estimates of the standard deviation. If both effects operate, the negative caseload-arbitrariness correlation could reflect study-level examiner pool size rather than the routinization interpretation I offer in the main text.

I test this directly. If the mechanical story is correct, then controlling for the number of examiners should attenuate the caseload coefficient, and the number of examiners should itself predict arbitrariness. Table 30 reports the results. For $|\beta^{\text{std}}|$, the caseload coefficient moves from -0.012 (SE $= 0.005$, $p < 0.01$) in the univariate specification to -0.014 (SE $= 0.005$, $p < 0.05$) after controlling for log(n_examiners). The coefficient becomes slightly larger in absolute value, not smaller. For the P10–P90 range, the point estimate follows the same pattern, moving from

-0.035 (SE $= 0.024$, $p = 0.15$) to -0.048 (SE $= 0.026$, $p = 0.07$) after the control is added, though neither range coefficient reaches conventional significance. In both models, the number of examiners is itself insignificant ($p = 0.45$ and $p = 0.21$, respectively).

The reason the univariate estimate is attenuated rather than inflated is that caseload per examiner and number of examiners are negatively correlated in the sample ($r = -0.44$): studies with fewer examiners tend to have higher caseloads per examiner. This means the mechanical confound, to the extent it exists, works in the opposite direction from the one that would threaten my interpretation. The negative caseload-arbitrariness relationship is substantive, not an artifact of examiner pool size. Appendix D.9 closes the two remaining mechanical readings: the relationship strengthens rather than weakens in the subsamples where the weak-instrument gate cannot bind, and a calibration simulation shows the noise correction cannot manufacture it.

Table 30: Caseload and arbitrariness: controlling for number of examiners.

	$\lvert\beta^{\text{std}}\rvert$		Range P10–P90	
	(1)	(2)	(3)	(4)
$\log_{10}$(caseload/yr)	-0.012^{**}	-0.014^{*}	-0.035	$-0.048^{\cdot}$
	(0.005)	(0.005)	(0.024)	(0.026)
log(n_examiners)		-0.003		-0.017
		(0.003)		(0.013)
N	83	83	90	90

Notes: Standard errors clustered by paper in parentheses. $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.10$. Columns (1) and (3) are univariate; columns (2) and (4) add log(n_examiners + 1) as a control. The caseload coefficient strengthens after controlling for examiner pool size, ruling out a mechanical relationship.

D.19 Is the rise at rare margins mechanical?

Are rare decisions more arbitrary than common ones, for the people who receive them? Figure 12 in the main text shows that the share of rare-decision recipients whose outcome a redraw would change is highest where decisions are rarest. That share is absolute disagreement divided by the base rate, so an objection could be: is the rise toward the rare end not mechanical, the same disputed cases spread over a smaller pile? Not on its own. If bureaucrats disagreed on a fixed share of all cases, each keeping a habitual grant rate, the share would rise twentyfold between a 40 percent margin and a 2 percent one, and would pass 100 percent just above a 4 percent base rate: more rare decisions changed than are made. I observe margins down to 2 percent, so that version of a purely mechanical reading does not hold in this sample; in Figure 47 the faint dotted curve carrying that reading leaves the figure.

Another more serious benchmark holds disagreement fixed in standards instead. Every bureaucrat applies the same bar, placed with the same spread; where the decision is rare, the

bar sits among extreme cases, few cases sit near it, and absolute disagreement shrinks almost in step with the base rate. That world produces only a 1.7- to 2.5-fold rise (the dashed curve in Figure 47; 2.5 under a probit model of the hidden scale, 1.7 under a logit one). Both references are drawn through the fitted data curve at the median base rate, so they differ only in how they extend toward rare margins. The data rise three- to four-fold: more than a shared bar allows, less than habitual shares would require.

So the answer is actually yes for this dataset: rare margins have more disagreement than equal standards can explain, though two cautions apply. The gap runs two to four standard errors, depending on the scale held fixed. The level is the same no matter the reading, though: at a 2 percent margin, about two in five recipients owe the outcome to the draw.

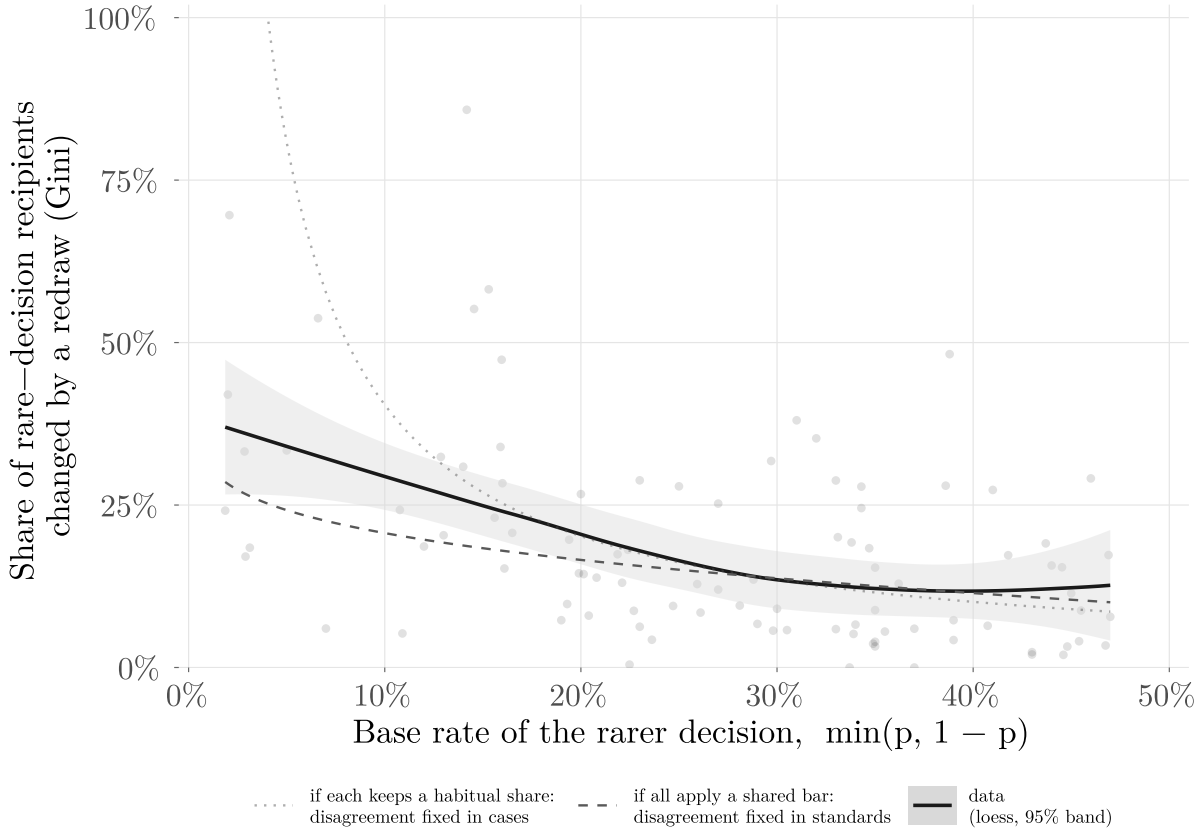


Figure 47: Each faint point is one study: the share of rare-decision recipients whose outcome a redraw of the bureaucrat would change (the Gini coefficient of the corrected decision-rate distribution), against the base rate of the rarer decision. The solid curve is a loess fit with its 95% band. The two reference curves are drawn through the fitted curve at the median base rate, so they differ only in shape; at the rarest margins the shared-bar reference falls below the band. The faint dotted curve fixes disagreement as a share of all cases, as when each bureaucrat keeps a habitual grant rate: the share rises twentyfold and passes 100 percent just above a 4 percent base rate, which would mean more decisions changed than were made, so this reading is impossible at the rarest observed margins. The dashed curve fixes disagreement in standards, a shared bar with a fixed spread on a hidden severity scale: absolute disagreement then shrinks almost in step with the base rate and the share rises about 2.5-fold (1.7 under a logit version). The data rise three- to four-fold.

D.20 Leniency IV studies, measures, and sources

Table 31: Overview of leniency IV studies, per-estimate measures, and input provenance. Each row is one main estimate. Measures are the corrected GMD, the lenient-decile versus strict-decile gap, and the rare-side redraw share. Superscript b marks a lower bound. Source pages identify inputs checked against the paper; other inputs are carried from the recorded extraction.

Study	Treatment and place	GMD / decile / rare-side	Inputs
Galasso and Schankerman (2014)	patent invalidated vs. patent upheld *US (Federal)*	– / – / –	*Not computable:* insufficient published inputs
Alexeev and Weatherburn (2022)	fine vs. other sanction *Australia*	0.133 / 0.301 / 0.200	leave-out rate (p. 6); SD 0.121; base 0.67; caseload (constr.) 280; J 273 (p. 4)
Araujo et al. (2023)	bankruptcy declared vs. not declared *Brazil*	0.107 / 0.242 / 0.267	leave-out rate (p. 13); SD 0.120; base 0.20; caseload 29; J 340; min 10
Agan et al. (2023)	case nonprosecuted vs. prosecuted *US (Massachusetts)*	0.033 / 0.074 / 0.080	leave-out rate (p. 15); SD 0.040; base 0.20; caseload 212 (p. 16); J 315 (p. 16); min 30 (p. 12)
Diaz and Salas (2022)	detained pretrial vs released pending trial *Colombia (Bogotá)*	0.243 / 0.559 / 0.858	leave-out rate; SD 0.222; base 0.86 (p. 6); caseload 67; J 112 (p. 9)
Humlum et al. (2025)	classroom training assigned vs. not assigned *Denmark*	0.033 / 0.075 / 0.042	leave-out rate (p. 17); SD 0.040; base 0.39; caseload (constr.) 318 (p. 56); J 577 (p. 56); min 50 (p. 56)
Berggren (2022)	antibiotic prescribed vs. not prescribed *Sweden*	0.076 / 0.173 / 0.197	leave-out rate (p. 16); SD 0.090; base 0.19
Aneja and Avenancio-León (2020)	prison vs. probation *US (Texas)*	0.057 / 0.129 / 0.073	leave-out rate (p. 35); SD 0.051 (p. 58); base 0.39; caseload 3506; J 37
Aizer and Doyle (2015)	juvenile incarceration vs. no incarceration *US (Illinois)*	0.040 / 0.090 / 0.087	leave-out rate (p. 775); SD 0.039 (p. 776); base 0.23; caseload 607 (p. 775); J 62 (p. 775); min 10 (p. 767)
McDonough et al. (2022)	detention vs. release *US (Pennsylvania)*	0.039 / 0.089 / 0.055	leave-out rate (p. 7); SD 0.035; base 0.35; caseload 16804; J 6 (p. 7)
Bald et al. (2022) *Estimate 1: Investigations sample, young boys (investigated before age 6)*	child removal vs. family preservation *US (Rhode Island)*	0.058 / 0.131 / 0.145	leave-out rate (p. 1932); SD 0.055 (p. 1934); base 0.20; caseload (constr.) 387 (p. 1934); J 102 (p. 1934); min 10 (p. 1946)
Bald et al. (2022) *Estimate 2: Investigations sample, young girls (investigated before age 6)*	child removal vs. family preservation *US (Rhode Island)*	0.058 / 0.131 / 0.138	leave-out rate (p. 1932); SD 0.055 (p. 1934); base 0.21; caseload (constr.) 387 (p. 1934); J 102 (p. 1934); min 10 (p. 1946)
Arbour and Marchand (2022)	parole granted vs. parole denied *Canada (Quebec)*	0.080 / 0.181 / 0.088	leave-out rate (p. 12); SD 0.074 (p. 14); base 0.46; caseload 518 (p. 6); J 20 (p. 6)
White (2019)	incarceration vs. no incarceration *US (Texas)*	0.032 / 0.089 / 0.034	full-sample rate (p. 318); SD 0.034 (p. 318); base 0.53; caseload 1512 (p. 314); J 75 (p. 314)

Continued on next page

Table 31 continued

Study	Treatment and place	GMD / decile / rare-side	Inputs
Arnold et al. (2018) *Estimate 1: Miami-Dade County (weekend bail hearings), 2006-2014*	release vs. detention *US (Florida)*	0.035 / 0.079 / 0.057	leave-out rate (p. 18); SD 0.033; base 0.70; caseload 1448 (p. 19); J 177 (p. 19)
Arnold et al. (2018) *Estimate 2: Philadelphia, 2010-2014*	release vs. detention *US (Pennsylvania)*	0.035 / 0.079 / 0.057	leave-out rate (p. 18); SD 0.033; base 0.70; caseload 1448 (p. 19); J 177 (p. 19)
Arteaga (2023)	parental incarceration vs. no incarceration *Colombia*	0.000 / 0.000 / 0.000	leave-out rate (p. 18); SD 0.043 (p. 68); base 0.37; caseload (constr.) 70 (p. 36); J 262 (p. 36); min 15 (p. 9)
Augustine et al. (2022)	diverted vs. standard prosecution *US (California)*	0.049 / 0.111 / 0.152	leave-out rate (p. 689); SD 0.044; base 0.16
Bakx et al. (2020)	care eligibility granted vs. denied *Netherlands*	0.068 / 0.155 / 0.207	leave-out rate (p. 5); SD 0.070 (p. 8); base 0.83; caseload 112 (p. 5); J 455 (p. 8); min 50 (p. 4)
Klaauw and Vethaak (2022)	job search requirement imposed vs. not imposed *Netherlands*	0.136 / 0.294 / 0.252	leave-out rate (p. 17); SD 0.151 (p. 22); base 0.73; caseload 92; J 461; min 50 (p. 17)
Hyman (2022)	TAA certified vs. not certified *US*	0.035 / 0.080 / 0.058	leave-out rate (p. 16); SD 0.040 (p. 41); base 0.69; caseload (constr.) 336; J 250 (p. 39); min 10 (p. 16)
Sampat and Williams (2019)	patent granted vs. patent rejected *US*	0.076 / 0.173 / 0.174	leave-out rate (p. 17); SD 0.090; base 0.22; min 10 (p. 17)
Black et al. (2024)	disability benefits granted vs. denied *US*	0.108 / 0.245 / 0.339	leave-out rate (p. 6); SD 0.096; base 0.84; caseload (constr.) 1922 (p. 7); J 1436 (p. 5); min 200 (p. 6)
Bos et al. (2023)	mental illness diagnosed vs. not diagnosed *Sweden*	0.019 / 0.043 / 0.332	leave-out rate (p. 13); SD 0.017; base 0.03; caseload 3923 (p. 11); J 102 (p. 11); min 500 (p. 11)
Cahuc et al. (2024)	positive compensation (wrongful dismissal) vs. none *France*	0.072 / 0.163 / 0.231	leave-out rate (p. 16); SD 0.066; base 0.84; caseload (constr.) 450 (p. 13); J 159 (p. 13); min 50 (p. 13)
Loeffler (2013)	incarceration vs. no incarceration *US (Illinois)*	0.041 / 0.093 / 0.059	full-sample rate (p. 149); SD 0.040; base 0.35; caseload 812; J 25
Gupta (2023)	patent granted vs. patent rejected *US*	– / – / –	*Not computable:* no dispersion statistic or examiner count reported anywhere incl. appendices
Dippel et al. (2025)	title transferred vs. not transferred *US*	0.226 / 0.513 / 0.353[b]	fitted values (p. 58); SD 0.200; base 0.32; caseload (constr.) 12; J 450
Honigsberg and Jacob (2021)	expungement granted vs. denied *US*	– / – / –	*Not computable:* list-mean instrument over 10-30 arbitrators with region-mean imputation; single-examiner noise model inapplicable
Herbst (2023)	IDR enrolled vs. not enrolled *US*	0.000 / 0.000 / 0.000	leave-out rate (p. 9); SD 0.025; base 0.34; caseload (constr.) 245 (p. 9); J 203 (p. 9); min 100 (p. 7)

Continued on next page

Table 31 continued

Study	Treatment and place	GMD / decile / rare-side	Inputs
Nagin and Snodgrass (2013)	incarceration vs. no incarceration *US (Pennsylvania (Centre, Crawford, Cumberland, Dauphin, Erie, Mercer counties))*	0.162 / 0.369 / 0.173	full-sample rate (p. 617); SD 0.147 (p. 617); base 0.53; caseload 255 (p. 608); J 24 (p. 617); min 100 (p. 608)
Autor et al. (2019)	disability benefits granted vs. denied *Norway*	0.053 / 0.120 / 0.203	leave-out rate (p. 2620); SD 0.050; base 0.13; caseload (constr.) 375 (p. 2620); J 75 (p. 2620); min 10 (p. 2622)
Chan et al. (2022)	antibiotic prescribed vs. not prescribed *US*	0.008 / 0.019 / 0.060	leave-out rate (p. 753); SD 0.010 (p. 98); base 0.07; caseload 1458 (p. 76); J 3199 (p. 748); min 100 (p. 747)
Chan et al. (2023)	high-priority classification vs. low-priority *US*	0.039 / 0.089 / 0.059	leave-out rate (p. 14); SD 0.035; base 0.33; caseload (constr.) 7255; J 1217 (p. 40); min 20 (p. 11)
Harding et al. (2018)	prison vs. probation *US (Michigan)*	0.017 / 0.039 / 0.027	leave-out rate; SD 0.022; base 0.31; caseload 884; J 124 (p. 86); min 100 (p. 62)
Harding et al. (2019)	prison vs. probation *US (Michigan)*	0.236 / 0.536 / 0.380	leave-out rate; SD 0.210; base 0.31; caseload 513; J 124; min 100
Munroe and Wilse-Samson (2013)	foreclosure completed vs. foreclosure dismissed *US (Illinois)*	0.374 / 0.850 / 0.482	leave-out rate (p. 16); SD 0.332; base 0.61; caseload 11722 (p. 40); J 12 (p. 8)
Bonfim and Nogueira (2024)	reorganization granted vs. denied *Portugal*	0.200 / 0.455 / 0.244	leave-out rate (p. 18); SD 0.190; base 0.59; caseload (constr.) 52; J 129 (p. 17)
Bonfim and Nogueira (2024)	reorganization granted vs. denied *Portugal*	0.224 / 0.509 / 0.273	leave-out rate (p. 15); SD 0.210; base 0.59; caseload (constr.) 52; J 129
Dominguez and Lazcano (2025)	pretrial detention vs. no pretrial detention *Chile*	0.052 / 0.119 / 0.242	leave-out rate (p. 10); SD 0.049; base 0.89; caseload 372; J 827; min 50 (p. 8)
Green and Winik (2010)	prison vs. probation *US (District of Columbia)*	0.144 / 0.369 / 0.173	full-sample rate (p. 370); SD 0.141; base 0.42; caseload 111; J 9
Baron and Gross (2025)	child removal vs. family preservation *US (Michigan)*	0.017 / 0.039 / 0.420	leave-out rate (p. 18); SD 0.025; base 0.02; caseload (constr.) 235 (p. 18); J 3011 (p. 18); min 50 (p. 18)
Eichmeyer and Zhang (2022)	opioid prescribed vs. not prescribed *US*	0.044 / 0.100 / 0.085	leave-out rate (p. 12); SD 0.045; base 0.26; caseload (constr.) 369; J 5313 (p. 8); min 200 (p. 8)
Eichmeyer and Zhang (2023)	opioid prescribed vs. not prescribed *US*	0.009 / 0.020 / 0.241	leave-out rate (p. 4); SD 0.014 (p. 6); base 0.02; caseload 133 (p. 3); J 4941 (p. 3); min 50 (p. 3)
Cohen (2024)	housing program enrolled vs. not enrolled *US (California)*	0.071 / 0.161 / 0.537	leave-out rate (p. 17); SD 0.070 (p. 18); base 0.07; caseload 65 (p. 80); J 236 (p. 80); min 31 (p. 80)
Emanuel et al. (2025)	involuntary hospitalization vs. release *US (Pennsylvania (Allegheny County / Pittsburgh))*	0.081 / 0.184 / 0.181	leave-out rate (p. 16); SD 0.098; base 0.78; caseload 39 (p. 17); J 424 (p. 17); min 15 (p. 18)
Blæhr and Søgaard (2021)	psychotherapy prescribed vs. not prescribed *Denmark*	– / – / –	*Not computable:* assignment to hospital departments not clearly random (KE 2026-08-26); distribution has a 79 percent point mass at zero

Continued on next page

Table 31 continued

Study	Treatment and place	GMD / decile / rare-side	Inputs
Leslie and Pope (2017) *Estimate 1: Felonies (IVNM / Lasso-selected instruments; preferred specification)*	detention vs. release *US (New York)*	0.072 / 0.163 / 0.084	leave-out rate (p. 10); SD 0.065; base 0.43; caseload (constr.) 1523 (p. 6); J 212 (p. 6); min 500 (p. 6)
Leslie and Pope (2017) *Estimate 2: Misdemeanors (IVNM / Lasso-selected instruments; interpret with caution due to sorting and disposition-at-arraignment issues)*	detention vs. release *US (New York)*	0.046 / 0.106 / 0.194	leave-out rate (p. 10); SD 0.042; base 0.12; caseload (constr.) 1523 (p. 6); J 212 (p. 6); min 500 (p. 6)
Enamorado et al. (2024)	detention vs. release *US (Florida)*	0.132 / 0.301 / 0.288	leave-out rate (p. 12); SD 0.120; base 0.23; caseload (constr.) 289; J 156 (p. 12); min 2 (p. 36)
French and Song (2014)	disability benefits granted vs. denied *US*	0.073 / 0.166 / 0.158	leave-out rate (p. 10); SD 0.066; base 0.77; caseload 1189 (p. 12); J 1497 (p. 12); min 50 (p. 12)
Helland and Yun (2023)	discovery conference held vs. not held *US (California)*	0.161 / 0.365 / 0.552	leave-out rate (p. 144); SD 0.143 (author email); base 0.15 (author email); caseload 387; J 75 (p. 142); min 100 (p. 142)
Gonçalves and Mello (2025)	harsh fine vs. lenient fine *US (Florida)*	0.191 / 0.434 / 0.278	leave-out rate (p. 11); SD 0.170; base 0.66; caseload 864 (p. 9); J 1960 (p. 9); min 50 (p. 9)
Lopes et al. (2022)	eligible vs not eligible for supported housing *Netherlands*	0.087 / 0.196 / 0.309	leave-out rate (p. 121); SD 0.088; base 0.86; caseload (constr.) 64; J 198; min 30
Rateb (2022)	release vs. detention *US (Illinois)*	0.039 / 0.088 / 0.067	leave-out rate (p. 23); SD 0.035; base 0.29; caseload 12794; J 12 (p. 12)
Dahl et al. (2014)	disability benefits granted vs. denied *Norway*	0.065 / 0.148 / 0.271	leave-out rate (p. 1721); SD 0.060; base 0.12; caseload (constr.) 380; J 79; min 10
Grindaker et al. (2021)	bankruptcy declared vs. not declared *Norway*	0.130 / 0.296 / 0.192	leave-out rate (p. 16); SD 0.125; base 0.66; caseload (constr.) 96 (p. 16); J 879 (p. 16); min 20 (p. 20)
Gupta et al. (2016)	release vs. detention *US (Pennsylvania)*	– / – / –	*Not computable:* instrument in log-bail units (leave-out mean of log bail); no propensity-scale dispersion reported
Fraisse (2017)	bankruptcy granted vs. denied *France*	0.138 / 0.314 / 0.157	leave-out rate (p. 700); SD 0.137; base 0.44; caseload 65 (p. 690); J 1296 (p. 690); min 10 (p. 690)
Wermink et al. (2024)	incarceration vs. no incarceration *Netherlands*	0.151 / 0.343 / 0.474	leave-out rate (p. 328); SD 0.143 (author email); base 0.16; caseload 56 (p. 327); J 735 (p. 325)
Adams-Prassl et al. (2024)	prison vs. probation *Finland*	0.011 / 0.026 / 0.052	leave-out rate (p. 16); SD 0.025; base 0.11; caseload (constr.) 186; J 308 (p. 53); min 100 (p. 9)

Continued on next page

Table 31 continued

Study	Treatment and place	GMD / decile / rare-side	Inputs
Ivanov and Wang (2024)	rating changed vs. rating unchanged *US*	0.190 / 0.431 / 0.288	leave-out rate (p. 12); SD 0.230; base 0.33; caseload 9 (p. 36); min 5 (p. 13)
Cespedes et al. (2025)	Chapter 13 granted vs. dismissed *US*	– / – / –	*Not computable:* insufficient published inputs
Williams and Weatherburn (2022)	electronic monitoring vs. prison *Australia*	0.083 / 0.189 / 0.324	leave-out rate (p. 3); SD 0.080; base 0.13; caseload (constr.) 119; J 139 (p. 7); min 10 (p. 6)
Gillooly (2022)	high-priority classification vs. low-priority *US*	0.038 / 0.090 / 0.098	leave-out rate (p. 10); SD 0.037; base 0.19; caseload 670; J 31 (p. 11)
Farre-Mensa et al. (2016)	patent granted vs. patent rejected *US*	0.168 / 0.397 / 0.245	training-sample rate (p. 14); SD 0.131; base 0.66; caseload (constr.) 771 (p. 15); min 10 (p. 14)
Humphries et al. (2025) *Estimate 1: Main 2SLS sample; margin c vs d (noncarceral conviction vs dismissal), instrumented by judge noncarceral-conviction stringency Zc controlling for incarceration stringency Zi*	incarceration vs. no incarceration *US (Virginia)*	0.034 / 0.077 / 0.057	leave-out rate (p. 22); SD 0.039; base 0.30; min 100 (p. 22)
Humphries et al. (2025) *Estimate 2: Main 2SLS sample; margin i vs c (incarceration vs noncarceral conviction), instrumented by judge incarceration stringency Zi controlling for dismissal stringency Zd*	incarceration vs. no incarceration *US (Virginia)*	0.037 / 0.083 / 0.040	leave-out rate (p. 22); SD 0.042; base 0.55; min 100 (p. 22)
Jordan et al. (2025) *Estimate 1: Cook County (Chicago) felony cases, 1990–2007; first offenders*	incarceration vs. no incarceration *US (Illinois)*	0.028 / 0.063 / 0.073	leave-out rate (p. 9); SD 0.028; base 0.19; caseload 842 (p. 35); J 44 (p. 35); min 500 (p. 72)

Continued on next page

Table 31 continued

Study	Treatment and place	GMD / decile / rare-side	Inputs
Jordan et al. (2025) *Estimate 2: Cook County (Chicago) felony cases, 1990–2007; repeat offenders*	incarceration vs. no incarceration *US (Illinois)*	0.045 / 0.102 / 0.066	leave-out rate (p. 9); SD 0.043; base 0.66; caseload 842 (p. 35); J 44 (p. 35); min 500 (p. 72)
Jordan and Kim (2025)	affidavit acquired vs. not acquired *US (Illinois)*	0.178 / 0.404 / 0.582	leave-out rate (p. 13); SD 0.160; base 0.85; caseload 186 (p. 8); J 11 (p. 8); min 50 (p. 8)
Doyle (2007) *Estimate 1: Cook County delinquency sample (first investigations July 1990–June 2001; outcome window through 2000; Medicaid-enrolled prior to report; school-age children 5–15 at investigation)*	child removal vs. family preservation *US (Illinois)*	0.065 / 0.147 / 0.120	leave-out rate (p. 1593); SD 0.092 (p. 1594); base 0.27; caseload (constr.) 38 (p. 1593); J 409 (p. 1593); min 10 (p. 1593)
Doyle (2007) *Estimate 2: Statewide employment sample (first investigations July 1990–June 2001; outcomes: employ-ment/earnings in 2002–H1 2003; Medicaid-enrolled prior to report; school-age children 5–15 at investigation)*	child removal vs. family preservation *US (Illinois)*	0.029 / 0.066 / 0.063	leave-out rate (p. 1593); SD 0.073 (p. 1594); base 0.23; caseload (constr.) 38 (p. 1593); J 409 (p. 1593); min 10 (p. 1593)
Doyle Jr. (2008)	child removal vs. family preservation *US (Illinois)*	0.091 / 0.206 / 0.284	leave-out rate (p. 9); SD 0.091; base 0.16; caseload (constr.) 74 (p. 9); J 733 (p. 9); min 10 (p. 9)
González-Uribe and Reyes (2021)	accelerator selected vs. not selected *Colombia*	0.067 / 0.151 / 0.128[b]	fitted values (p. 16); SD 0.059; base 0.26; caseload 8 (p. 10); J 50 (p. 11)
Kissel and Roy (2025)	C-section vs. vaginal delivery *US*	0.058 / 0.132 / 0.144	leave-out rate (p. 35); SD 0.078; base 0.20; caseload (constr.) 47; J 12256 (p. 75); min 100 (p. 13)
Kleinberg et al. (2018)	release vs. detention *US (New York)*	0.084 / 0.190 / 0.158[b]	grouped rates (p. 261); base 0.74
Lee (2023)	residential housing assigned vs. direct release *US (Iowa)*	0.062 / 0.140 / 0.088	leave-out rate (p. 11); SD 0.070; base 0.35; caseload 120; J 61; min 50 (p. 8)
Lum et al. (2017)	bail set vs. no bail *US (New York)*	– / – / –	*Not computable:* near-far matched-pair design; no judge count or judge-level propensity distribution reported
Machado Ribeiro (2019)	detention vs. release *Brazil (Rio de Janeiro)*	0.127 / 0.289 / 0.184	leave-out rate (p. 25); SD 0.116 (p. 26); base 0.65; caseload 318 (p. 17); J 18 (p. 17)

Continued on next page

Table 31 continued

Study	Treatment and place	GMD / decile / rare-side	Inputs
Maestas et al. (2013)	disability benefits granted vs. denied *US*	0.108 / 0.244 / 0.154	leave-out rate (p. 14); SD 0.100; base 0.65; caseload 250 (p. 10); J 8700 (p. 10); min 30 (p. 9)
Bhuller et al. (2018)	incarceration vs. no incarceration *Norway*	0.073 / 0.166 / 0.075	leave-out rate (p. 5); SD 0.072; base 0.52; caseload (constr.) 238; J 597; min 50
Bhuller et al. (2020)	incarceration vs. no incarceration *Norway*	0.083 / 0.189 / 0.085	leave-out rate (p. 1278); SD 0.080; base 0.51; caseload (constr.) 258 (p. 1278); J 500 (p. 1278); min 50 (p. 1282)
Bhuller and Sigstad (2022)	incarceration vs. no incarceration *Norway*	0.073 / 0.167 / 0.078	leave-out rate (p. 6); SD 0.072; base 0.53; caseload (constr.) 241; J 596; min 50
Grimon (2026)	child removal vs. family preservation *US (Pennsylvania)*	0.137 / 0.312 / 0.154	leave-out rate (p. 12); SD 0.137; base 0.45; caseload 61 (p. 45); J 519 (p. 45); min 10 (p. 60)
Famulari (2012)	remedial class required vs. not required *US (California)*	0.216 / 0.491 / 0.280	leave-out rate (p. 9); SD 0.229; base 0.39; caseload (constr.) 15 (p. 11); J 81; min 5 (p. 11)
LaForest-Tucker (2026) *Estimate 1: Board member stringency (estimated from Figure 3 histogram in user-provided image)*	parole granted vs. denied *US (Pennsylvania)*	0.017 / 0.039 / 0.020	leave-out rate (p. 13); SD 0.015; base 0.57; caseload (constr.) 28302; J 9
LaForest-Tucker (2026) *Estimate 2: Hearing examiner stringency (estimated from Figure 3 histogram in user-provided image)*	parole granted vs. denied *US (Pennsylvania)*	0.020 / 0.046 / 0.024	leave-out rate (p. 13); SD 0.018; base 0.57; caseload (constr.) 12736; J 20
Yin et al. (2023)	VR services received vs. not received *US*	0.167 / 0.379 / 0.191	leave-out rate (p. 179); SD 0.168; base 0.44; caseload (constr.) 39; J 70; min 10
Grau et al. (2023)	detention vs. release *Chile*	0.032 / 0.075 / 0.112	leave-out rate (p. 14); SD 0.041; base 0.14; caseload (constr.) 367 (p. 14); J 907 (p. 14)
Norris et al. (2021)	incarceration vs. no incarceration *US (Ohio)*	0.054 / 0.123 / 0.090	leave-out rate; SD 0.049; base 0.30 (p. 2940); caseload 3129; J 256; min 100
Eren and Mocan (2021)	juvenile incarceration vs. no incarceration *US (Louisiana)*	0.047 / 0.106 / 0.095	leave-out rate (p. 39); SD 0.050 (p. 38); base 0.25; caseload (constr.) 238 (p. 36); J 73 (p. 38); min 25 (p. 36)
Horsch et al. (2021)	patent invalidation vs. no invalidation *US*	0.103 / 0.233 / 0.114	leave-out rate (p. 8); SD 0.110; base 0.45; caseload 65
Nesbit (2023)	mental health treatment mandated vs. not mandated *US (North Carolina)*	0.033 / 0.076 / 0.334	leave-out rate (p. 22); SD 0.030; base 0.05; caseload (constr.) 1992; J 365; min 10 (p. 99)
Di Tella and Schargrodsky (2013)	detention vs. release *Argentina*	0.029 / 0.048 / 0.696	leave-out rate (p. 47); SD 0.033 (p. 46); base 0.02; caseload (constr.) 116 (p. 59); J 206 (p. 58); min 10 (p. 59)

Continued on next page

Table 31 continued

Study	Treatment and place	GMD / decile / rare-side	Inputs
Hjalmarsson and Lindquist (2019)	military service vs. no military service *Sweden*	0.139 / 0.354 / 0.279	leave-out rate (p. 2533); SD 0.130 (p. 2536); base 0.75; caseload (constr.) 3456; J 67 (p. 2534); min 100 (p. 2531)
Diamond et al. (2020)	foreclosure completed vs. foreclosure dismissed *US (Illinois)*	0.026 / 0.058 / 0.037	leave-out rate (p. 17); SD 0.024; base 0.35; caseload (constr.) 3600; J 77
Collinson et al. (2024) *Estimate 1: Cook County, IL (2000–2016 court records; linked labor-market sample used for first stage)*	eviction ordered vs. no eviction *US (Illinois)*	0.028 / 0.063 / 0.040	leave-out rate (p. 25); SD 0.027 (p. 26); base 0.65; caseload (constr.) 1600 (p. 26); J 127 (p. 26); min 100 (p. 9)
Collinson et al. (2024) *Estimate 2: New York City, NY (2007–2016 court records; linked labor-market sample used for first stage)*	eviction ordered vs. no eviction *US (New York)*	0.023 / 0.052 / 0.032	leave-out rate (p. 25); SD 0.023 (p. 26); base 0.35; caseload (constr.) 1600 (p. 26); J 127 (p. 26); min 100 (p. 9)
Gong et al. (2025)	patent grant vs. no grant *US*	0.189 / 0.429 / 0.318	full-sample rate (p. 15); SD 0.170; base 0.70 (p. 7)
Helénsdotter (2026)	child removal vs. family preservation *Sweden*	0.045 / 0.083 / 0.186	leave-out rate (p. 17); SD 0.059; base 0.88; caseload (constr.) 82; J 249; min 25 (p. 17)
Antill (2022)	reorganization vs. liquidation *US*	– / – / –	*Not computable:* judge count and direct SD absent from published paper (SD 0.195 itself inferred from coefficients); online appendix not on disk
Axbard (2024)	conviction vs. acquittal *Philippines*	0.268 / 0.560 / 0.291	leave-out rate (p. 12); SD 0.238; base 0.46; caseload 26; J 41 (p. 11); min 12 (p. 12)
Mullainathan and Obermeyer (2021)	diagnostic testing ordered vs. not ordered *US*	0.010 / 0.022 / 0.171	leave-out rate (p. 704); SD 0.023 (p. 705); base 0.03; caseload (constr.) 62 (p. 706); J 3951 (p. 706)
Bernstein et al. (2019)	reorganization vs. liquidation *US*	0.052 / 0.119 / 0.064	leave-out rate (p. 24); SD 0.061 (p. 25); base 0.41
Markussen et al. (2012)	graded (partial) vs. non-graded (full) sick-leave certificate *Norway*	0.058 / 0.131 / 0.130	training-sample rate (p. 22); SD 0.075 (p. 18); base 0.22; caseload (constr.) 57 (p. 18); J 3868 (p. 18)
Amaral et al. (2023)	arrest vs. no arrest *UK*	0.012 / 0.026 / 0.184	leave-out rate (p. 11); SD 0.012; base 0.03; min 400 (p. 12)
Didwania (2020)	release vs. detention *US (Federal)*	0.044 / 0.100 / 0.060	leave-out rate (p. 38); SD 0.050; base 0.37; caseload 241 (p. 35); J 452 (p. 30); min 20 (p. 35)
Koppel et al. (2024)	pretrial detention vs. release at arraignment *US (New York City)*	0.078 / 0.177 / 0.135	leave-out rate (p. 6); SD 0.070; base 0.29; caseload 1273 (p. 9); J 63 (p. 4); min 500 (p. 4)
Caskurlu (2019)	infringement found vs. no infringement *US (Federal)*	0.158 / 0.359 / 0.295[b]	fitted values (p. 15); SD 0.140; base 0.73; min 4 (p. 13)

Continued on next page

Table 31 continued

Study	Treatment and place	GMD / decile / rare-side	Inputs
Cortés (2025) *Estimate 1: Any juvenile incarceration (public-attorney quality IV; court-by-year residualized leave-out incarceration rate)*	detention vs. release *Chile*	0.020 / 0.046 / 0.043	leave-out rate (p. 14); SD 0.043 (p. 15); base 0.24; caseload (constr.) 118 (p. 9); J 545 (p. 15); min 10 (p. 14)
Cortés (2025) *Estimate 2: Pretrial detention (detention-judge leniency IV; court-by-year residualized leave-out leniency)*	detention vs. release *Chile*	0.002 / 0.005 / 0.004	leave-out rate (p. 14); SD 0.020 (p. 15); base 0.22; caseload (constr.) 118 (p. 9); J 545 (p. 15); min 10 (p. 14)
Dobbie and Song (2015)	Chapter 13 granted vs. dismissed *US*	0.033 / 0.074 / 0.037	leave-out rate (p. 15); SD 0.030 (p. 15); base 0.45; caseload 4377 (p. 9); J 112 (p. 9); min 10 (p. 9)
Dobbie et al. (2018)	detention vs. release *US (Florida)*	0.029 / 0.053 / 0.032	leave-out rate (p. 16); SD 0.030; base 0.55; caseload 2380 (p. 17); J 179 (p. 17)
Dobbie et al. (2017)	Chapter 13 granted vs. dismissed *US*	0.017 / 0.039 / 0.019	leave-out rate (p. 861); SD 0.024 (p. 860); base 0.45; caseload (constr.) 729; J 348 (p. 857); min 10 (p. 857)
Grönqvist et al. (2024)	parental incarceration vs. no incarceration *Sweden*	0.054 / 0.122 / 0.095	leave-out rate (p. 13); SD 0.051; base 0.28; caseload (constr.) 582; J 604 (p. 11); min 50 (p. 11)
Dobbie et al. (2021)	loan take-up vs. no loan *UK*	0.035 / 0.080 / 0.052	leave-out rate (p. 22); SD 0.047 (p. 22); base 0.66; caseload 180 (p. 21); J 254 (p. 21); min 50 (p. 21)
Ofek-Shanny et al. (2025)	care eligibility granted vs. denied *Israel*	0.093 / 0.209 / 0.129	leave-out rate (p. 10); SD 0.090; base 0.64; caseload (constr.) 171 (p. 8); J 299 (p. 8); min 100 (p. 8)









Appendix E

Paper 2 Appendices

E.1 Theoretical Elaborations

I. Distinction: Bias and Noise

To further elaborate on the concepts of bias and noise and their orthogonal relationship, presented in Figure 1, consider the analogy of a symphony orchestra aiming to perfectly execute a composer's score. First, a performance characterized by low bias and low noise (Figure 1A) would be flawless: every note is in tune, rhythms are precise, and all musicians are perfectly synchronized, faithfully rendering the composer's intent. This epitomizes consistently accurate judgment.

Second, an orchestra exhibiting high bias but low noise (Figure 1B) might, for instance, consistently play a quarter-tone flat or ten beats per minute too slowly. All musicians perform cohesively in this incorrect manner (low noise), yet the overall performance systematically deviates from the score (high bias). This mirrors a scenario wherein decision-makers reliably agree, but on an erroneous outcome.

Third, a performance marked by low bias but high noise (Figure 1C) would be discordant. Individual musicians commit distinct errors—some playing sharp, some flat, some rushing, and others dragging. While the overall intended melody might still be discernible (low average bias), such individual inconsistencies mar the listening experience. This is analogous to decision-makers whose judgments, on average, may be correct, but whose individual assessments are highly variable and unreliable.

Finally, high bias and high noise (Figure 1D) would culminate in a chaotic and off-target performance. The orchestra might, on average, play too rapidly (high bias), while individual musicians are also erratic with pitch and timing (high noise). This represents the worst-case scenario, where judgments are both systematically incorrect and highly variable.

This orchestral analogy, much like the visual representation in Figure 1 and the recruitment example, underscores that bias concerns the average deviation of judgments from the "true" target (the score's accurate and correct interpretation), whereas noise pertains to the scatter or variability among individual judgments (the musicians' individual performances).

II. Connection: Noise and Legitimacy

Noise in public sector judgment and decision-making contravenes the legitimacy of the public sector institutions. This erosion affects both 'normative legitimacy' (the alignment of bureaucratic authorities and their operations with principled democratic-administrative norms and values; Beetham, 2013; Peters, 2018) and 'perceived legitimacy' (the public's view of these entities as appropriate, competent, and fair; Suchman, 1995; Tyler, 2006; Fine Licht et al., 2022). When

pivotal decisions affecting citizens' lives are subject to the caprice of random error rather than the principled application of rules, the foundation of rational-legal authority is weakened (Weber, 1978). Such inconsistency undermines the predictability of public institutions and, by extension, challenges the principle of legal certainty. Consequently, public trust in public agencies—an invaluable asset for democratic governance (Behn, 2001; Bouckaert 2012; Pierre and Peters, 2020; Van de Walle and Bouckaert, 2003; Van Ryzin, 2011) that influences policy compliance, service engagement, and democratic participation (Chanley et al., 2000; Lee and Schachter, 2019; Tyler and Jackson, 2014)—is potentially eroded. This erosion stems not merely from public awareness of inconsistency but from the failure to deliver consistently fair and rule-bound governance.

III. Concerns: AI Deployment in Government

Prominent concerns relating to technical, ethical, and democratic risks associated with the integration of AI in the public sector are as follows:

a) The risk of embedding or amplifying biases present in training data, leading to disparate impacts and discriminatory outcomes (Barocas and Selbst, 2016; Danks and London, 2017; Eubanks, 2018; Köchling and Wehner, 2020; Sunstein, 2022)—aptly termed "weapons of math destruction" (O'Neil, 2016);

b) The opacity and lack of explainability in 'black box' AI tools (Fine Licht and Fine Licht, 2020; Grimmelikhuijsen, 2023; Janssen et al., 2022)—although similar critiques regarding cognitive inscrutability have been leveled against human decision-making itself (Bonezzi et al., 2022)

c) Difficulties in assigning accountability when AI systems err and the potential for "moral crumple zones" (Elish, 2019)

d) The potential erosion of human expertise (Balasubramanian et al., 2022; Fügener et al., 2021).

Moreover, certain demographic and social groups exhibit marked skepticism toward the use of AI in the public sector (Humlum and Vestergaard, 2025; Sidhu et al., 2024)—often attributed to cognitive factors such as 'motivated reasoning' (Hemesath and Tepe, 2025) or 'algorithm aversion' (Burton et al., 2020; Dietvorst et al., 2015; Mahmud et al., 2022).

IV. Preference: Specialization over Ensembling

The preference may arise because specialization aligns more readily with existing workflows, preserves a greater degree of human autonomy and discretion— factors known to influence technology acceptance (Dietvorst et al., 2018; Ingrams et al., 2022)—and may encounter less organizational resistance than modes that accord AI a more direct and empowered role. Such considerations are particularly salient given the current political climate and prevailing skepticism toward comprehensive AI adoption in high-stakes public decisions (Kuziemski and Misuraca, 2020). This reluctance resonates with broader 'algorithm aversion' (Burton et al., 2020; Mahmud

et al., 2022) and concerns among parts of the public toward public sector AI integration (Hemesath and Tepe, 2025; Humlum and Vestergaard, 2025; Sidhu et al., 2024).

E.2 Sample Characteristics

As shown in Table A1, the sample comprised predominantly female (57.2%), middle-aged (53.2% aged 40-59), and highly educated (79.3% held a university degree) professionals. Participants also reported extensive experience in applicant assessment, with a substantial majority (78.0%) performing this task at least monthly: 14.8% did so daily, 45.5% multiple times per week or month, and 17.7% monthly. Conversely, 14.8% of respondents reported undertaking this task 'almost never.'

Table A1. Descriptive Sample Statistics

	Primary Sample (OS2 "AI advice")	**Robustness Sample** (OS2 full sample)
Gender	57.2%	55.6%
- Female	32.8%	35.5%
- Male	10.0%	08.9%
- Diverse/No answer		
Age	12.2%	11.7%
- 20-29	23.2%	21.4%
- 30-39	26.6%	27.7%
- 40-49	26.6%	25.7%
- 50-59	05.2%	07.3%
- 60-69	06.3%	06.3%
- No answer		
Education	07.7%	09.9%
- Apprenticeship, vocational training in the dual system	07.0%	07.1%
- Vocational school degree	23.2%	22.1%
- Bachelor	14.4%	14.5%
- Master	33.6%	33.1%
- Diploma	08.1%	06.9%
- PhD	05.9%	06.5%
- No answer		
Task expertise: Applicant assessment	14.8%	13.6%
- Almost never	17.7%	17.3%
- Once a month	27.7%	28.3%
- Several times a month	04.1%	03.9%
- Once a week	13.7%	15.6%
- Several times a week	01.5%	01.1%
- Once a day	13.3%	12.8%
- Several times a day	07.4%	07.4%
- No answer		
N	271	538

Notes: Distribution of sample demographics and task expertise. The task expertise variable reflects respondents' self-reported frequency of assessing job candidates.

E.3 Derivation of the 'Specialization-Ensemble Weighting Point' (SEWP)

The derivation of the SEWP rests on a simplifying assumption: that the behavioral process of specialization can be modeled as if each rater, on average, assigns a consistent weight (α) to the AI advice. Under this model, a rater's post-advice rating ($Post_i$) for a given candidate is a weighted average of their pre-advice rating (Pre_i) and the constant AI advice ($Advice$):

$Post_i = (1 - \alpha)Pre_i + \alpha(Advice)$ (Eq. A1)

Because the AI $Advice$ is a constant value for any given candidate profile across all raters, its variance is zero. Consequently, the variance of the post-advice ratings, $Var(Post)$, is determined solely by the variance of the pre-advice ratings and the weight applied to them. Applying the properties of variance to Equation A1:

$Var(Post) = Var((1 - \alpha)Pre + \alpha(Advice))$ (Eq. A2)

Given that $Var(\alpha(Advice)) = 0$, this simplifies to:

$Var(Post) = (1 - \alpha)^2 Var(Pre)$ (Eq. A3)

Taking the square root of both sides of Equation A3 yields the relationship between the standard deviations:

$SD_{post} = (1 - \alpha)SD_{pre}$ (Eq. A4)

Rearranging Equation A4 to solve for α gives us the formula for the SEWP, which represents the average equivalent AI weight that explains the observed reduction in variance:

$\alpha_{SEWP} = 1 - \frac{SD_{post}}{SD_{pre}}$ (Eq. A5)

E.4 Results of Levene's Test

Table A2. Levene's Test (Homogeneity of Variances)

	Specialization	**Ensembling**				
	(post advice)	10% AI	20% AI	30% AI	40% AI	50% AI
Candidate Profile						
#1	F=9.64**	F=2.15	F=9.51**	F=23.54***	F=45.86***	F=77.96***
#2	F=12.96***	F=2.84	F=12.55***	F=31.08***	F=60.53***	F=102.89***
#3	F=5.17*	F=1.82	F=8.04**	F=19.91***	F=38.78***	F=65.93***
#4	F=4.71*	F=2.58	F=11.40***	F=28.23***	F=54.99***	F=93.48***
Overall	F=24.43***	F=10.13**	F=43.91***	F=102.41***	F=173.29***	F=271.62***

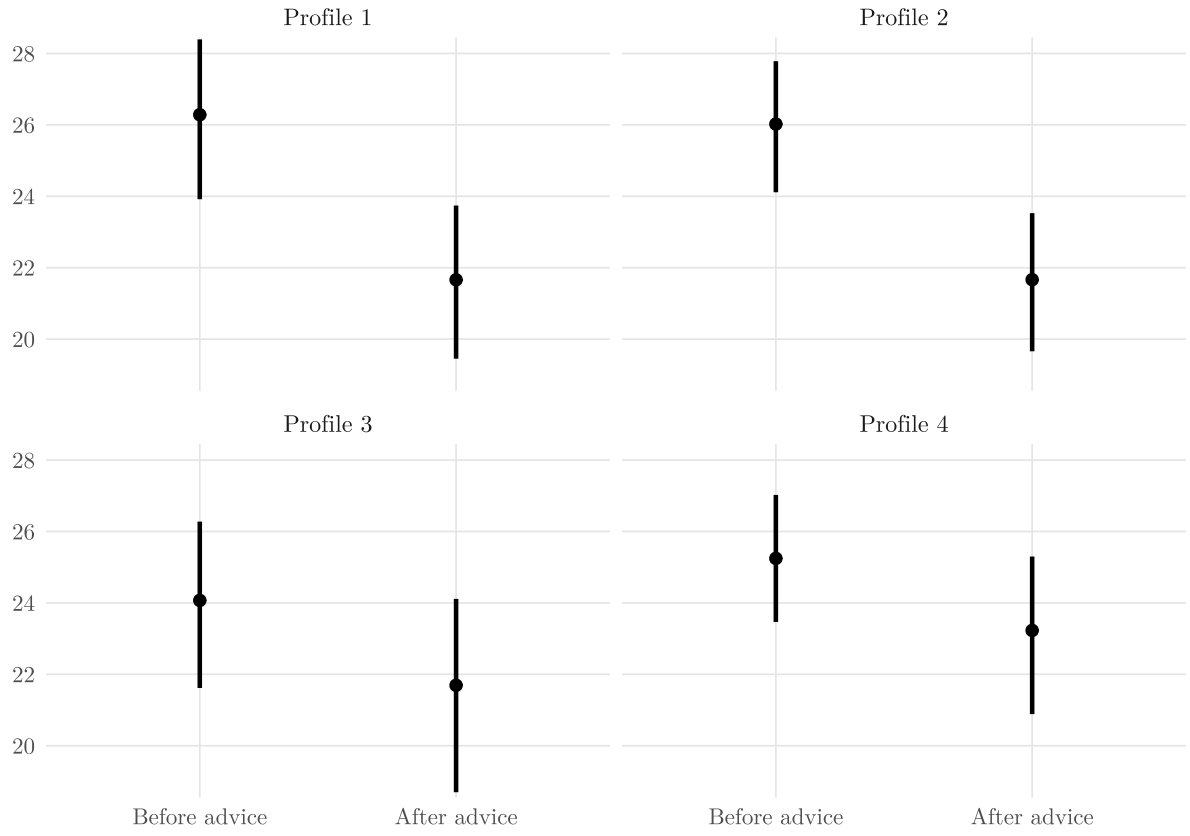
Notes: * $p < .05$; ** $p < .01$; *** $p < .001$. H_0: equal variances. Results of tests against the variance of the human-only pre-advice ratings.

E.5 Robustness Analysis by Candidate Profiles

In the *specialization* mode, the noise-reduction effect is robust across all four candidate profiles. As illustrated in Figure A1, the introduction of AI advice led to reductions in the standard

deviation of ratings for each profile (Levene's test, $p < .05$ for all; see Appendix E.4).

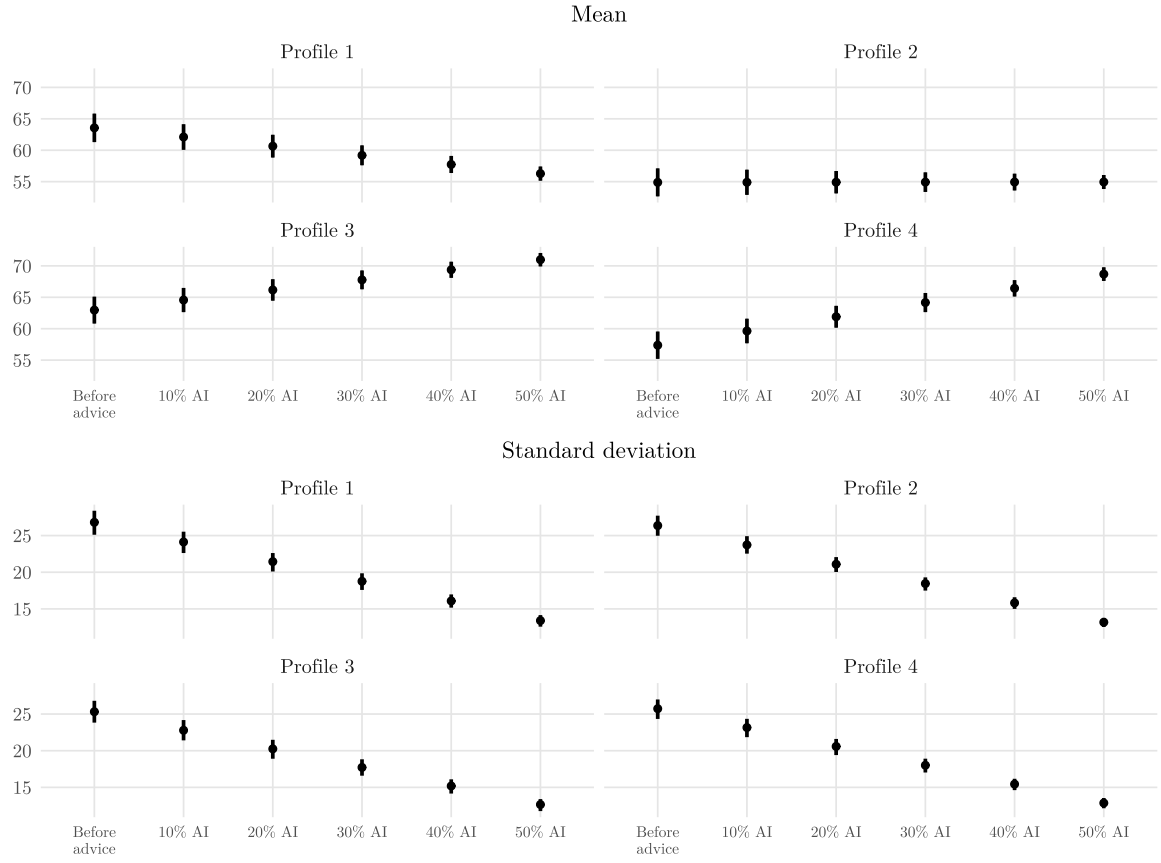
Figure A1. Standard Deviations for Specialization by Candidate Profiles



Notes: This figure provides a disaggregated, profile-level view of the specialization analysis presented in the main text (Figure 3, bottom left panel).

In the *ensembling mode*, the dose-response relationship between increasing AI weight and decreasing noise persists across all four candidate profiles (Figure A2). Although the noise reduction is not statistically significant when the AI's weight is only 10%, it is highly significant and substantial for all four profiles at AI weights of 20% and above (Levene's test, $p < .001$ for all; see Appendix E.4).

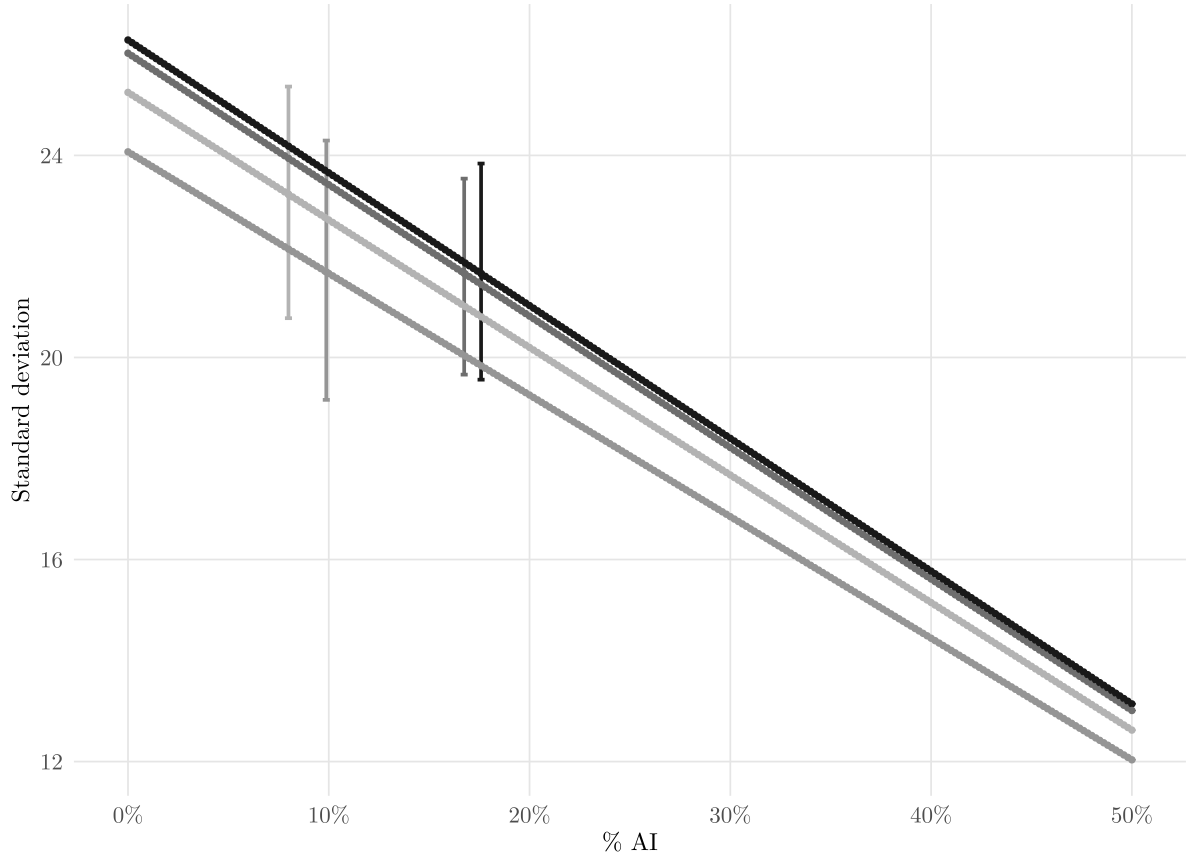
Figure A2. Mean Job Fit Ratings and Standard Deviations for Ensembling by Candidate Profiles



Notes: This figure provides a disaggregated, profile-level view of the ensembling analysis presented in the main text (Figure 3, right panel).

Moreover, we estimate the *SEWP* for each candidate profile individually to examine the consistency of our main finding. As shown in Figure A3, this analysis confirms the robustness of our aggregate result, while also revealing expected variance, with the profile-level SEWP estimates corresponding directly to the magnitude of noise reduction observed for each candidate. The SEWP ranged from 17.6% for the profile with the greatest noise reduction to 8.0% for the one with the least. Importantly, all profile-level SEWP estimates are positive and substantively meaningful, indicating that the specialization mode consistently reduces noise across all four cases.

Figure A3. Comparing the Noise-Reduction Efficacy of Specialization and Ensembling by Candidate Profiles

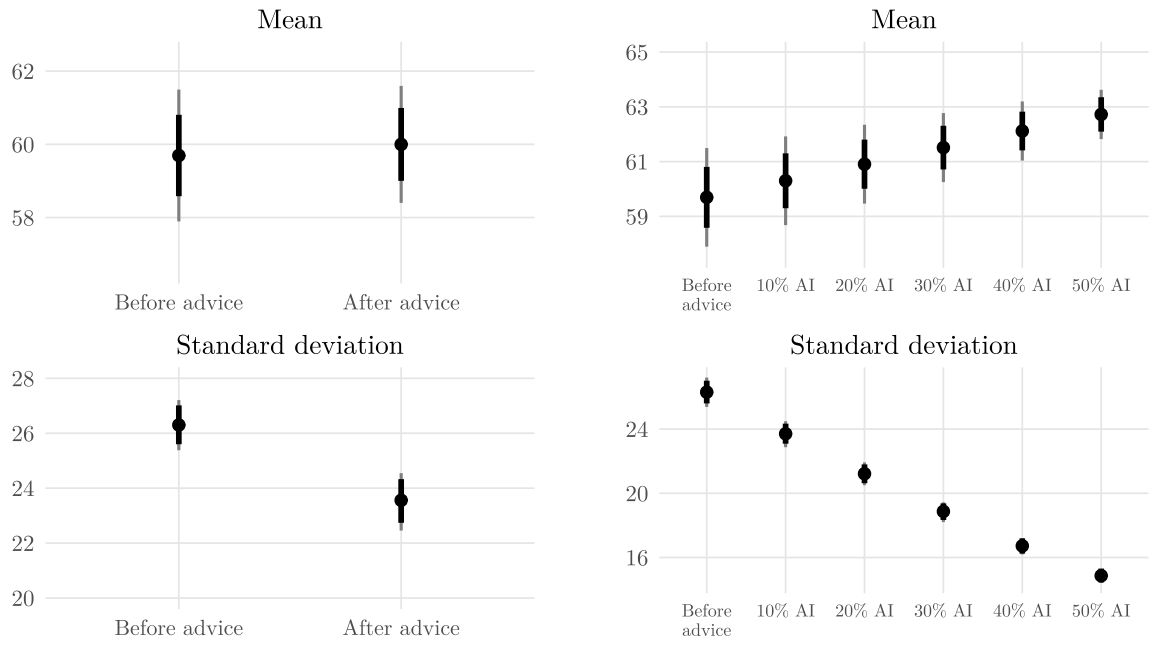


Notes: This figure disaggregates the analysis presented in Figure 4 of the main text to illustrate the noise-reduction efficacy for each of the four candidate profiles individually. Each candidate profile is represented by a distinct color gradient.

E.6 Robustness Analysis Using the Full Sample

As shown in Figure A4, both the specialization and ensembling modes demonstrate substantial and significant noise-reduction effects in the full sample analysis.

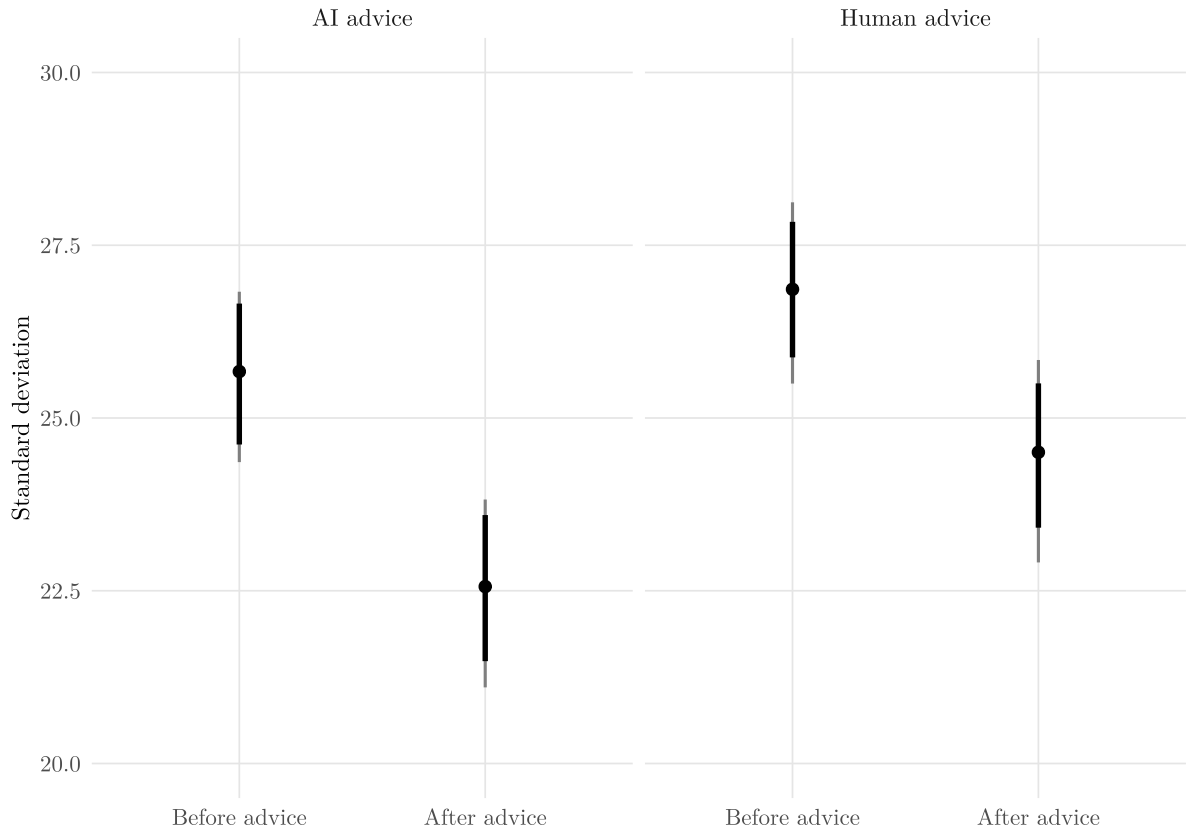
Figure A4. Mean Job Fit Ratings and Standard Deviations by Decision Mode, Full OS2 Sample



Notes: This figure replicates the analysis presented in Figure 3 of the main text but is based on the full OS2 sample.

Moreover, the full sample analysis allows us to explore whether the noise-reduction efficacy of specialization depends on the perceived source of the advice. As illustrated in Figure A5, although AI-framed advice produced a slightly larger point-estimate reduction in standard deviation, this difference is not statistically significant ($p > .05$). This null result suggests that specialization's noise-reduction effect may be driven by a general anchoring mechanism, whereby judgments converge on a common reference point irrespective of its perceived source (AI versus human peer). In terms of implications, this null finding provides no clear evidence that recruiters in our sample were systematically more or less receptive to advice from an AI compared to a human peer for reducing judgmental noise. It suggests that even if individual attitudes like 'algorithm aversion' (Burton et al., 2020) or 'algorithm appreciation' (Logg et al., 2019) were present, they did not translate into a measurable difference in the aggregate noise-damping effect.

Figure A5. Standard Deviations for Specialization by Source of Advice



Notes: This figure compares the noise-reduction efficacy of AI-framed versus human-framed advice within the specialization mode. Each panel displays the standard deviation of job fit ratings before and after participants received advice. The left panel shows the effect for advice framed as originating from "an AI," while the right panel shows the effect for advice framed as from "a human peer." Vertical lines represent two types of bootstrapped 95% confidence intervals: unclustered (thicker lines) and clustered at the respondent level (thinner lines). Clustering accounts for the non-independence of ratings provided by the same participant.

E.7 Methodological Limitations

Given the nature of the data, our findings are best understood as proof-of-concept and their extrapolation should be inferred in light of the following considerations.

First, our analysis of ensembling was based on a post-hoc simulation rather than a live implementation. Future research should examine ensembling in more realistic settings where recruiters are aware of a predefined aggregation rule and its practical consequences (e.g., for interview selection), which would allow for the observation of any strategic behavioral responses.

Second, our specialization analysis used an assessment-information-reassessment design. Real-world specialization might involve concurrent rather than sequential exposure to AI advice, a temporal difference that could alter the magnitude of the observed noise reduction.

Third, our findings emerge from a specific context: a single public organization, one job position, four candidate profiles, and one proprietary AI. The reliance on a proprietary tool

raises concerns regarding the stability of results across software versions or different AI platforms. However, it is essential to distinguish between the AI's specific fit assessment and the AI's role as a low-variance anchor. The noise-reduction efficacy of ensembling and specialization depends on the AI signal remaining more consistent than the human judgment it seeks to augment. While the magnitude of noise reduction might vary across different algorithmic architectures, the direction and significance of the effect are theoretically robust as long as the AI provides a more stable assessment than the human rater. Nevertheless, the generalizability of our results warrants future validation across diverse platforms.

Similarly, the generalizability of our findings to administrative domains other than personnel selection and to other national contexts warrants further validation, particularly given the nuanced ways human-AI collaboration can shape judgment (cf. Glickman and Sharot, 2025). Similarly, our ensembling findings may be sensitive to the inherent properties and quality of the specific AI tool that we examine. Furthermore, future research should critically evaluate whether fully automated systems—algorithms and AI without human involvement—are superior to human-AI collaboration in reducing noise (Vaccaro et al., 2024). However, such an approach introduces distinct challenges related to bias, acceptance, ethics, and implementability, which must be weighed against the benefits of noise reduction.

Fourth, the experimental setting involved a survey-based evaluation. This artificial rating environment lacks the high stakes and rich information available in real-world recruitment decisions, where managers possess more comprehensive information and face greater consequences for their choices.

Finally, our primary dependent variable was decisional noise. A comprehensive evaluation must also assess the impact of human-AI collaboration modes on systematic bias and, ultimately, decision accuracy (e.g., the predictive validity of hiring decisions). Furthermore, the proprietary nature of the AI limits full transparency into its model architecture and training data—a common challenge in AI research that can impede a deeper understanding of the underlying mechanisms and limit exact scientific replication across different tools. A holistic research agenda should therefore aim to integrate measures of noise, bias, and overall error while prioritizing evaluations of open-source models to establish the long-term reliability of these findings.









Appendix F

Paper 3 Appendices

F.1 Further descriptives

Table 32: Distribution of case types

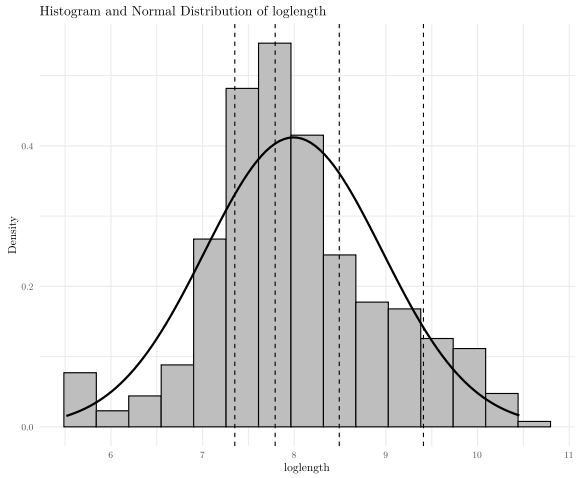
Case topic	Frequency
Hjælpemidler - §112	5848
Revurdering og forlængelse - § 24-29	5222
Tabt arbejdsfortjeneste §§ 42,43 + tilbagebetaling	4814
Uarbejdsdygtighed - § 7	4715
Merudgiftydelse §41 + tilbagebetaling	4696
Voksne - merudgifter - § 100	4292
Hjælp til enkeltudgifter - §§ 81-81a	4226
Hjælp til sygebehandling - §§ 82-82a	4049
Biler - §§ 114 og 115	3721
Tilbagebetaling af løbende hjælp - §§ 90 - 97, 69i	3469
Sanktioner - §§ 35-44a	3232
Børn - samvær - § 71	2911
Arbejdsgiverrefusion - øvrige betingelser § 54, 56, 57, 58 og 60	2716
Forbrugsgoder - §113	2676
Hjælp til flytning - § 85	2626
Fradrag for indtægter - §§ 30-33	2459
Hjemmehjælp § 83	2456
Børn - foranstaltninger - §§ 52 stk. 3 nr. 1-9, 55, 69 stk.1	2378
Formue - §§ 14 og 15	2353
Arbejdsgiverrefusion - anmeldelse § 59	2342
Bortfald af retten til sygedagpenge - §§ 21-23	2293
Voksne - socialpædagogisk bistand - § 85	2106
Social begivenhed - §11-12a + § 22	2094
Særlig støtte - § 34	1958
Boligindretning - § 116	1953
Personlig hjælp og ledsagelse §§44, 45 + tilbagebetaling	1872
Anmeldelse - §§ 35-39 og 41, 42 og 43	1857
Børn - økonomisk hjælp - § 52a	1693
Voksne - borgerstyret personlig assistance - § 96	1436
Beskæftigelseskrav - selvstændig / lønmodt - §§ 30-34 og 42	1338
Bevillingstidspunkt og satser - §§ 23-26 samt § 22 integrationsydelse	1338
Hjælp til visse persongrupper - §§ 27-29	1158
Forsørgerpligt, bopæl, udlandsophold mv. - §§ 2-5	1146
Ressourceforløbsydelse under ressourceforløb og sanktioner - §§ 68-69h	1102
Voksne - botilbud - §§ 107-110 (2013-2018)	1059
Ledighedsydelse §§ 74-77b	1017
Voksne - midlertidigt botilbud - § 107	845
Sygedagpenge - øvrige bestemmelser §§ 3 - 6, 8 - 20	790
Udnyttelse af arbejdsmuligheder - §§ 13,13C, 13g og 13f	789
Pasningstilbud, hjemmetræning m.m. §§32, 32a, 36, 39-40 + tilbagebet.	768
Voksne - ledsageordning - § 97	766
Børn - efterværn - §§ 68 og 76 (2013-2019)	725

Continued on next page

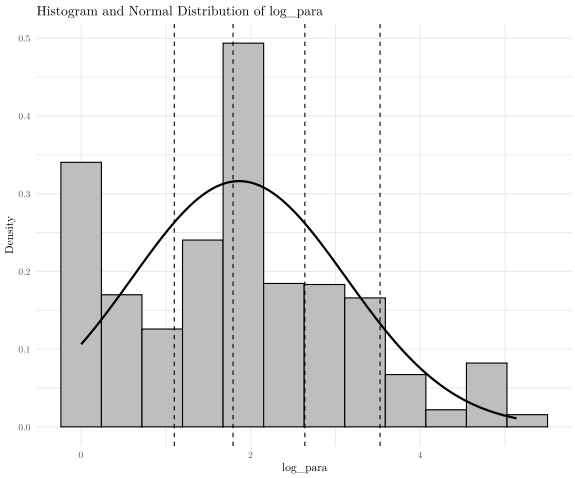
Paper 3: Distribution of case types (continued)

Case topic	Frequency
Tilbagebetaling § 71 og § 56-aftale	646
Voksne - kontante tilskud § 95	614
Bevillingstidspunkt og satser - §§ 22-26 (undtaget §§ 25 b-e)	599
Tilbagebetaling af hjælp efter- §§81-85a og §§ 91-97	598
Ressourceforløbsydelse under jobafklaring og sanktioner - §§ 69j-69u dog ikke §69s	597
Beregning af sygedagpenge - §§ 46-53	579
Børn - forældrebetaling §159	566
Børn - Efterværn og oprethold anbringelse - §§ 76 og 76a	467
Voksne - afløsning og aflastning - § 84	464
Børn - Børnefaglig undersøgelse - § 50	412
Voksne - behandling - §§ 101 og 102 (2013-2018)	400
Voksne - længerevarende botilbud - § 108	369
Voksne - behandling - § 102	362
Voksne - beskyttet beskæftigelse og aktivitets- og samværstilbud - §§ 103-106 (2013-2018)	347
Særlig hjælp vedrørende børn - § 83	330
Pasning af nærtstående - §§ 118-122	304
Børn - Foranstaltninger - §§ 52, stk. 3 og 55	295
Tilbagebetaling af løbende månedlig hjælp til forsørgelse - § 91	274
Voksne - genoptræning - § 86	272
Voksne - aktivitets- og samværstilbud - § 104	239
Børn - Forebyggende indsats - § 11	221
Voksne - misbrugsbehandling - § 101	201
Sanktioner §§ 36-38	198
Børn - økonomisk hjælp - § 11 (2013-2019)	189
Voksne - selvudpeget hjælp - §§ 94-94b	177
Børn - Valg af anbringelsessted - § 68b	171
225 timers reglen - § 13 f og 13 g	156
Børn - Hjemgivelse og hjemgivelsesperiode - § 68	151
Opholdskrav - § 11, stk. 3 (IY eller KH/UH)	138
Voksne - Krisecentre, forsorgshjem og herberg - §109 og §110	118
Indplacering på kontanthjælpsloftet - §§ 25 b-e	103
Husvilde - § 80	97
Voksne - støtte og kontaktperson - §§ 98 og 99	94
Voksne - ulovligt skolefravær, standsning af børne- og ungeydelse - § 155c	87
Voksne - beskyttet beskæftigelse - §§ 103 og 105, stk. 2	72
Voksne - rehabilitering - § 83a	72
Sanktioner § 39-40	71
Revalideringsydelse - §§ 70-73j	60
Børn - Forhold under anbringelsen - § 69, stk. 1	50
Sanktioner §§ 42-43	43
Forsørgelse af børn - § 84	34
Magtanvendelse, voksne - øvrige - §§ 124, 125, 128-128f og §§ 136a-137j	30
Magtanvendelse, voksne, fastholdelse mv. - § 126	21
Sanktioner § 40a	18
Børn - Bekendtgørelse om voksenansvar § 9	15
Vejledning og opfølgning - §§ 7-10	11
Magtanvendelse, voksne, alarmsystemer - § 125	8
Hjælp til efterlevende - § 85a	6
Magtanvendelse, voksne, besøgsrestriktioner - § 137b og 137c stk. 2	5
Magtanvendelse, voksne, indgreb i selvbestemmelsesretten, anbragte i boform v/strafferetlig afg. - § 137g-j	3
Hastigt progredierende lidelser SEL § 117 a	2
Magtanvendelse, voksne, beskyttelsesmidler, stofseler - § 128	2
Magtanvendelse, voksne, tilbageholdelse i boligen - § 127	1

In the following graphs the dotted lines represent the 25th, 50th, 75th, and 90th percentiles of the data.



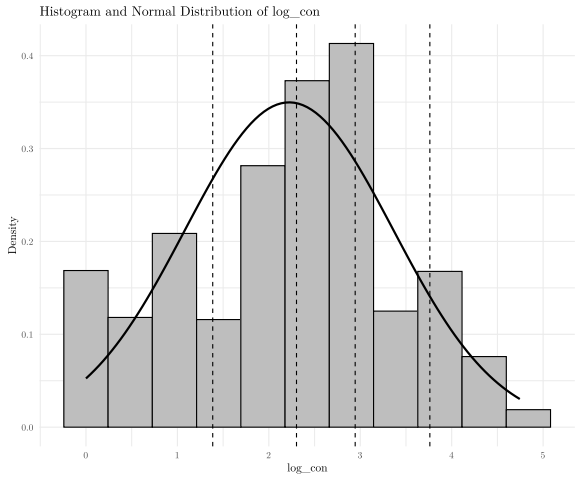
(a) Distribution of log(length)



(b) Distribution of log(cross-references)

Figure 48: Distribution of law characteristics

(a) Distribution of log(Conditionals)



(b) Distribution of log(Permissive Delegation)

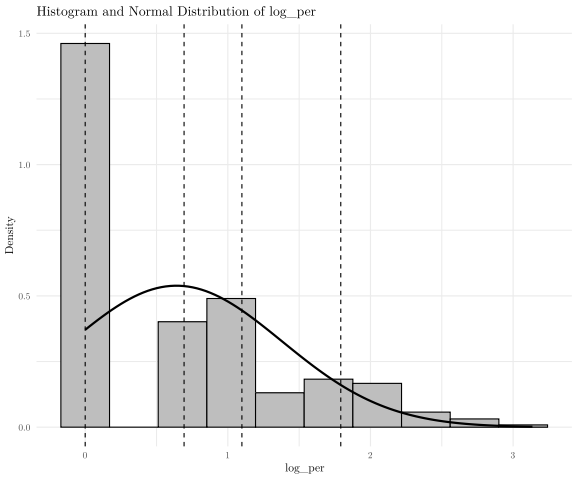
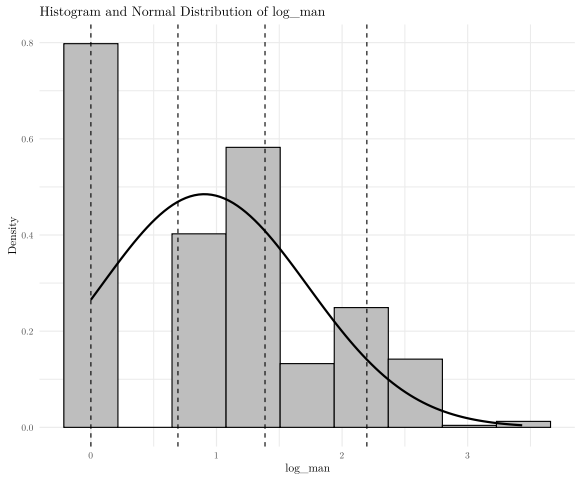


Figure 49: Distribution of law characteristics

(a) Distribution of log(Mandatory Delegation)



(b) Distribution of LIX-score

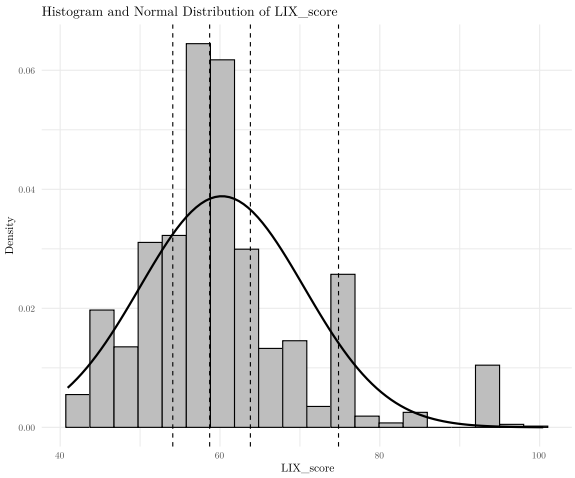
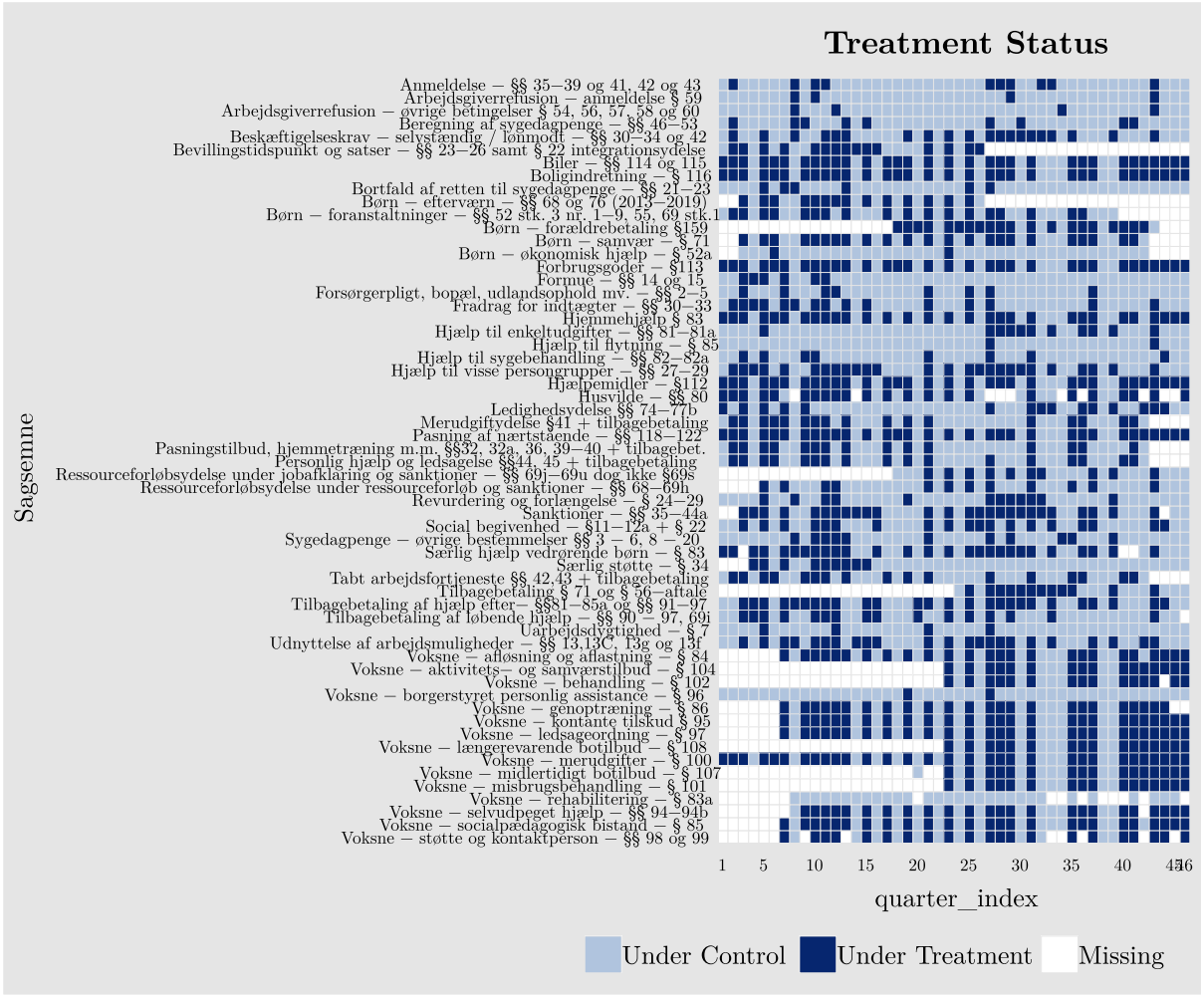


Figure 50: Distribution of law characteristics

F.2 Panel balance

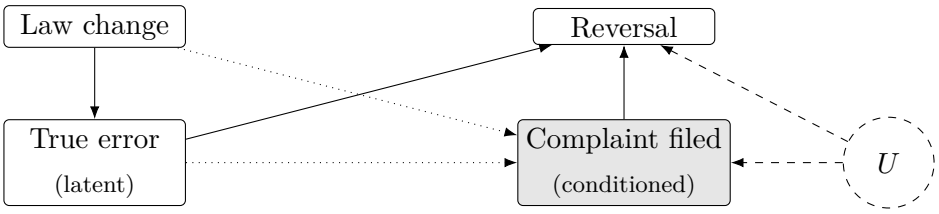
For the plot below, we have used the `panelView` package developed by Mou et al. (2023). The "under treatment" label simply refers to a change to the paragraph.

Figure 51: panelView



F.3 Is there collider bias?

For the paper, it is crucial that the law changes do not affect the propensity of citizens to file a complaint if the citizen partly chooses to complain based on the probability of there being a reversal. If legal changes also influence the number or the composition of citizens who complain, conditioning on complaints may create collider bias. This arises if (1) citizens' decision to complain depends partly on their perceived probability of reversal, and (2) true errors both increase the likelihood of reversal and make citizens more likely to complain. In that case, reforms that raise willingness to complain may bring more marginal cases into the system, thereby lowering the observed reversal rate. This is illustrated in the following DAG from the paper:



Conditioning on appealed cases can thus induce selection bias if reforms affect the propensity to file a complaint, and if unobserved case merits affect both the likelihood of reversal and the likelihood of complaining. If this is true, then an implication of this is that we should be able to assess whether reforms are associated with changes in complaint volume. A practical check is therefore to test whether our law-change measures predict complaint volume. We estimate the same fixed-effects specification with the number of complaints as the outcome. We find no meaningful association, suggesting that reforms do not systematically shift appeal behavior in a way that would mechanically drive our main results. Even if the correlation were significant, the point estimates are positive, which would then suggest that the marginal cases are not filing a complaint, which would suggest that the estimates are conservative as this would bias the error rate upwards.

	Length (log)	Cross-references (log)	Mandatory delegation	Permissive delegation	LIX score	Conditionals (log+1)
log(length)	−0.061					
	(0.119)					
log(paragraf_henvisning)		−0.041				
		(0.071)				
log(mandatory_delegation + 1)			0.029			
			(0.087)			
log(permissive_delegation + 1)				−0.069		
				(0.074)		
LIX_score					−0.020	
					(0.018)	
log(num_conditionals + 1)						0.030
						(0.057)
Num.Obs.	2851	2368	2849	2849	2851	2849
R2	0.815	0.791	0.815	0.816	0.816	0.815
R2 Adj.	0.806	0.779	0.806	0.806	0.806	0.806
RMSE	0.56	0.56	0.56	0.56	0.56	0.56
Std. errors	by: Sagsemne	by: Sagsemne	by: Sagsemne	by: Sagsemne	by: Sagsemne	by: Sagsemne
FE: Sagsemne	X	X	X	X	X	X
FE: DatePeriode	X	X	X	X	X	X

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
SEs clustered by case topic

F.4 Results measured with logit

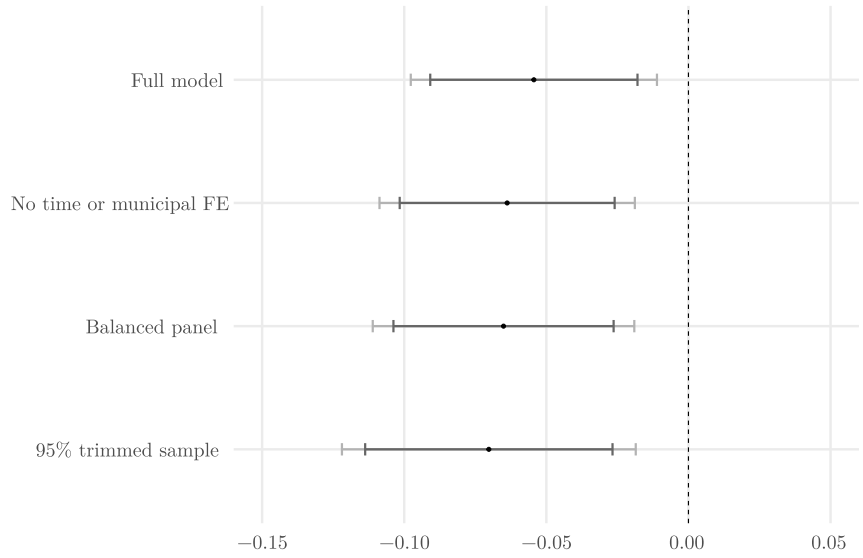


Figure 52: LOGIT: Dot-and-whisker plot of the coefficient estimates from the models using log(length)

	Full sample (logit)	95% trimmed sample (logit)	Balanced panel (logit)	No time FE (logit)
log(length)	−0.251*	−0.336*	−0.303*	−0.283*
	(0.110)	(0.138)	(0.120)	(0.111)
Num.Obs.	112 655	108 434	89 294	112 655
R2	0.044	0.046	0.044	0.038
R2 Adj.	0.041	0.043	0.041	0.037
RMSE	0.48	0.48	0.48	0.48
Std. errors	by: CaseTopic	by: CaseTopic	by: CaseTopic	by: CaseTopic
FE: CaseTopic	X	X	X	X
FE: Municipality	X	X	X	
FE: Time	X	X	X	

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 33: Regression results from the logit models

F.5 Results for other text-based measures

	Cross-references	Permissive delegation	Mandatory delegation	Readability (LIX)	Syntactic complexity (Adjectives)	Conditional clauses
log(paragraf_henvisning + 1)	−0.032*					
	(0.014)					
log(permissive_delegation + 1)		−0.024				
		(0.015)				
log(mandatory + 1)			−0.006			
			(0.013)			
LIX_score				−0.002		
				(0.003)		
log(num_adjectives + 1)					−0.064	
					(0.080)	
log(num_conditionals + 1)						−0.021
						(0.016)
Num.Obs.	112 663	112 566	112 566	112 663	112 566	112 566
R2	0.056	0.056	0.056	0.056	0.056	0.056
R2 Adj.	0.054	0.054	0.054	0.054	0.054	0.054
RMSE	0.48	0.48	0.48	0.48	0.48	0.48
Std. errors	by: CaseTopic	by: CaseTopic	by: CaseTopic	by: CaseTopic	by: CaseTopic	by: CaseTopic
FE: CaseTopic	X	X	X	X	X	X
FE: Municipality	X	X	X	X	X	X
FE: Time	X	X	X	X	X	X

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 34: Regression results from the auxiliary linear probability models

F.6 Precedent rulings and the timing of legal change

A natural concern is that legislators amend the provisions that are already error-prone, so that the negative association between length and the error rate reflects reform-chasing mistakes rather than detail-clarifying decisions. We can try to test this. When municipal practice on a topic is repeatedly overturned or in doubt, the National Social Appeals Board issues a *principmeddelelse*, a precedent ruling that fixes the binding interpretation for that class of cases. Precedent rulings are a visible, error-driven response by the same body that decides our outcome, so they show what selection on errors looks like in this setting.

In Column 2 of Table 35 we regress whether a topic receives a precedent ruling in a quarter on its lagged error rate, with topic and quarter fixed effects, and the coefficient is positive and significant (0.089, $p = 0.026$; a logit gives 1.42, $p = 0.005$). Column 1 uses the same regression for amendments, and there a topic's lagged error rate does not predict whether its statute changes the following quarter (-0.010, $p = 0.72$). This suggests that the Board's precedent rulings track

errors, whereas the statutory amendments do not.

	Amendment (1/0)	Precedent ruling (1/0)
	(1)	(2)
Lagged error rate	-0.0096	0.0889*
	(0.0270)	(0.0392)
Case-topic FE	Yes	Yes
Quarter FE	Yes	Yes
Observations (topic-quarters)	2,756	2,681
Clusters (case topics)	94	92

LPM; SEs clustered by case topic in parentheses. * $p < 0.05$.

Table 35: Do recent errors predict legal change or precedent rulings? Linear probability models; the outcome is an indicator for a statutory amendment (column 1) or a precedent ruling (column 2) on the topic that quarter. Standard errors clustered by case topic in parentheses. The same predictive design fires for precedent rulings but not for amendments.

We can also look at the numbers over time. Figure 53 plots the raw error rate in the quarters around a topic's first amendment. Error levels are flat across the eight quarters before reform, and a joint test of the pre-reform leads cannot reject zero ($F = 0.63$, $p = 0.73$).

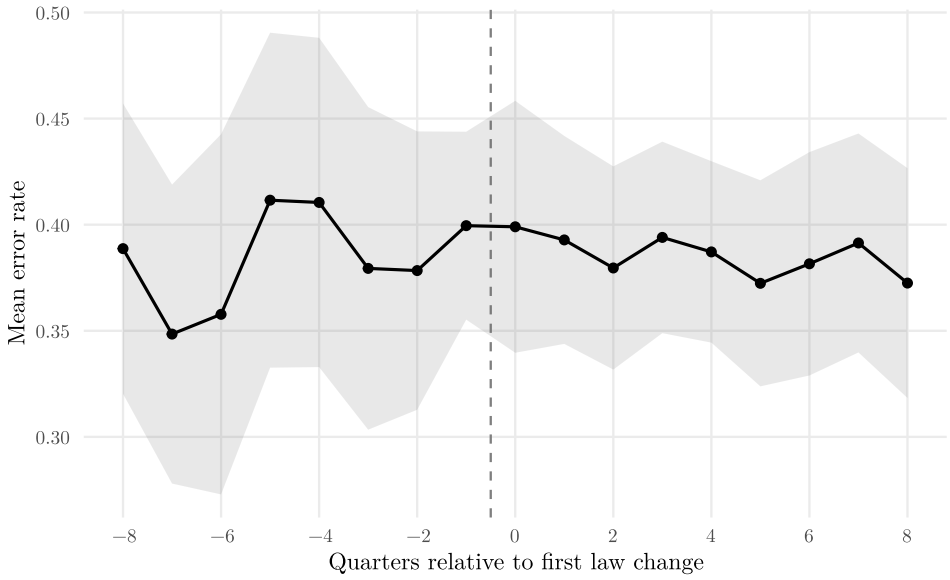


Figure 53: Error rate in raw levels around a topic's first law change. Flat pre-reform leads (joint $F = 0.63$, $p = 0.73$) indicate no selection on pre-trends.

F.7 Parliamentary debates: do legislators invoke the appeals system?

How much are politicians aware of errors? We can compare the specific debates that accompanied the changes to the laws we measure on versus debates on changes to other laws. These are rough measures, but they might be indicative. About 88% of the target debates mention implementation difficulty, but so do about 86% of all other debated Danish bills.

Restricting to explicit appeals-and-error vocabulary, focusing on actual mentions of the National Social Appeals Board, its precedent rulings, reversal and remand, and administrative-context error, about two-fifths of the target debates mention appeals and errors compared with a fifth of comparison bills. The National Social Appeals Board is named on the floor in about one in five of the target debates. The casework and appeals vocabulary is several times over-represented in the target debates relative to comparison bills, while generic complaint language is not (Table 37).

	Target bills (3 laws)	All comparison Danish bills	Social/empl. cmte. bills	Diff. vs all comparison	95% CI
(i) Broad implementation-difficulty incidence (debate-length detector)					
≥1 mention	88.4	85.7	89.4	+2.7	[-2.9, +8.2] n.s.
N (bills)	138	1872	226	vs cmte. −1.0	[-7.7, +5.7] n.s.
(ii) Narrow appeals/error vocabulary, excl. generic klage-					
≥1 mention	38.4	18.0	25.7	+20.4	[+12.1, +28.7]*
N (bills)	138	1872	226	vs cmte. +12.7	[+2.8, +22.7]*
(iii) National Social Appeals Board (Ankestyrelsen) named on the floor					
named by a speaker	19.6	2.1	–	+17.5	[+10.8, +24.1]*
N (bills)	138	1872	–		

Table 36: Parliamentary-speech base-rate comparison. Share of debated bills whose floor debates trip each measure, for the bills amending the three laws (Target) against comparison Danish bills. The broad implementation-difficulty measure does not separate the target bills from other bills; the narrow appeals-and-error vocabulary does. 95% confidence intervals in brackets; $^*p < 0.05$. The like-for-like comparison is for 2012–2022, the span of the comparison corpus.

Narrow-dictionary stem	Target bills (share of hits)	Comparison (share of hits)	Ratio (T / C)
Ankestyrelsen (Appeals Board)	31.7	6.2	5.1×
fejl in administrative context	39.4	14.8	2.7×
principafgørelse (precedent ruling)	1.3	0.3	4.1×
omgørelse (reversal)	1.3	0.8	1.7×
hjemvisning (remand)	0.0	0.3	0.0×
generic *klage-* (complaint)	26.3	77.6	0.3×
Total narrow hits	457	4023	

Table 37: Composition of the narrow appeals-and-error hits. Share of narrow-vocabulary hits accounted for by each stem, target against comparison bills. The casework and appeals vocabulary is over-represented in the target debates; generic complaint language (*klage*) is not.

F.8 The consultation record

We can also look at what is mentioned in the consultations preceding the changes in the laws. Like the parliamentary speeches, these are also mostly rough measures.

For many of the law changes, we have the ministry's notes (høringsnotat) and the stakeholders' responses (høringssvar) attached to the amending bills. These documents are not available for 21% of the changes during COVID-era fast-track bills.

The documents on these three laws are distinctively oriented toward the appeals system and citizens' legal certainty: relative to a base rate of other amending bills, the National Social Appeals Board is named about twice as often, legal certainty (*retssikkerhed*) appears about twice as often, individual-assessment language about five times as often, and calls for clarification about one and a half times as often (Table 38).

Generic "implementation" language is no more common in these documents than in amending bills generally, and explicit error-rate language, reversal or remand percentages, is essentially absent from both. Stakeholders and the ministry argue using comments on the appeals system and legal certainty.

Concept	Target rate	Base rate	Ratio	Welch p	Survives base rate?
Appeals system (*Ankestyrelsen*)	1.50	0.75	2.0×**	0.004	Yes
Legal certainty (*retssikkerhed*)	9.5%	4.0%	2.4×	–	Yes
Individual assessment	1.8%	0.4%	5.0×	–	Yes
Clarity / clarification	2.46	1.50	1.6×**	0.001	Yes
Implementation	1.90	1.78	1.1×	0.672	Generic
Discretion (*skøn*)	0.36	0.18	2.0×*	0.043	Fragile
Error / overturn	0.10	0.12	0.9×	0.789	No
Error rate (*omgørelsesprocent*)	0.00	0.06	0.0×	0.116	Absent

Social-policy ministry høringsnotater ($n = 102$) versus non-social-policy amending-bill notater ($n = 129$), 2013–2024, same keyword dictionary. For the first row *rate* is mean dictionary hits per 1,000 words; the *retssikkerhed* and individual-assessment rows report each stem's share of all dictionary hits (the comparison is robust to overall verbosity). Welch p-values are from a two-sample t-test on per-document intensity; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. "Survives base rate?": Yes = elevated and robust; Generic = present but indistinguishable from the base rate; Fragile = Welch-significant but the rank test is null ($p = 0.29$); No = not elevated; Absent = essentially zero in both corpora. Substantive (boilerplate-stripped) incidence, reported in the text, is 55%/17% for Ankestyrelsen, 41%/20% for *retssikkerhed*, and 20%/5% for individual assessment (social policy vs base); these are soft figures from a one-off filter. Raw mention-incidence of Ankestyrelsen is 80% vs 30%.

Table 38: Consultation-record register. Intensity (hits per 1,000 words) of each concept in the ministry notes and stakeholder responses on the three laws (Target) against a base rate of other amending bills, with the Welch test and whether the concept survives the base rate. The appeals system and legal-certainty register is distinctively elevated; implementation language is generic and explicit error-rate language is absent.

Concept raised by stakeholders	Bills (responses that raise it)	Notat engages it	Take-up rate
Appeals system	70	68	97%
Clarity / clarification	66	63	95%
Implementation	69	64	93%
Discretion (*skøn*)	43	25	58%
Error / overturn	31	17	55%

Bills with both a text-extractable høringsnotat and at least one text-extractable høringssvar ($n = 80$). For each concept, of the bills whose stakeholder responses raise it (dictionary intensity > 0), the share whose ministry notat also engages it. The ministry answers appeals, clarity and implementation concerns far more readily than concerns framed as error or as caseworker discretion.

Table 39: Ministry responsiveness. Among bills with both a note and consultation responses, the share for which the ministry engages a concern that stakeholders raise, by theme.

F.9 Classifying the content of paragraph changes

We classify all 345 length-changing amendments to the paragraphs in our window by content using LLMs and a keyword comparison of the old and new text rather than a full legal reading. Most changes are non-substantive: they add cross-references, procedural and documentation requirements, or clarified definitions, alongside a technical residual, and together these account for about 85% of changes.

Amendments that alter eligibility are few, 14.5% of changes, or 25.2% under a deliberately broad flag for any eligibility-related language.

Table 40: Classification of paragraph-length changes by content

Primary category	Count (of 345)	Share of changes (%)	Share of movement (%)
Cross-references to other provisions	83	24.1	29.4
Procedural requirements	56	16.2	31.3
Definitions and clarifications	6	1.7	1.5
Technical / residual edits	150	43.5	10.8
Substantive eligibility changes	50	14.5	27.0
Any eligibility marker (upper bound)	87	25.2	57.3

Notes: First-pass keyword classification of all 345 length-changing amendments (token-level diff against the previous quarter; ties broken toward the substantive category). "Share of changes" is the count as a percent of 345. "Share of movement" is the category's sum of $|\Delta \log(\text{length})|$ as a percent of the within-topic total. The substantive minority is larger by movement than by count: it is 14.5% of changes but 27.0% of the movement, and the upper-bound set is 25.2% of changes but 57.3% of the movement. The first five rows are mutually exclusive and sum to 345; the upper-bound row is a non-exclusive superset of the substantive row (any eligibility marker anywhere in the diff). The classification is keyword-based and the two-coder Cohen's κ audit has not yet been run.

If the negative length-error association came from reforms changing who is eligible, removing the eligibility-altering variation should pull the coefficient toward zero. But it does not; every estimate is at least as negative as the baseline -0.054. The drop-whole-topic variants grow in magnitude but lose statistical significance, which reflects a loss of power rather than attenuation.

Table 41: Robustness to excluding or neutralizing substantive changes

Specification	Coef. on log(length)		N	Clusters
	Estimate	(SE)	(obs.)	(topics)
Baseline (all 345 changes)	-0.054^{*}	(0.022)	112,663	95
Censor: change to next change	-0.060^{*}	(0.024)	99,637	95
Censor: ± 4 quarters	-0.059^{*}	(0.028)	96,603	95
Neutralize flagged variation	-0.069^{*}	(0.032)	112,663	95
Neutralize (upper bound)	-0.103	(0.055)	112,663	95
Drop primary-substantive topics	-0.092	(0.063)	66,171	72
Drop upper-bound topics	-0.095	(0.070)	56,757	65

Notes: Linear probability model of the reversal indicator on log(length), with municipality, case-topic, and quarter fixed effects; standard errors clustered by case topic in parentheses. $^{*}\,p < 0.05$. The Baseline row uses all 345 length-changing amendments. The remaining rows remove or neutralise the variation that the keyword classification (Table 40) flags as substantive: the censor variants drop topic-quarters around flagged changes, the neutralise variants zero the flagged log-length deltas and re-cumulate on the full sample, and the drop variants remove every affected case topic. No variant moves the coefficient toward zero; the drop variants lose significance through power loss (clusters fall from 95 to 72/65, and within-fixed-effect length variation falls by about a third), not attenuation.







Appendix G

Paper 4 Appendices

G.1 PRISMA

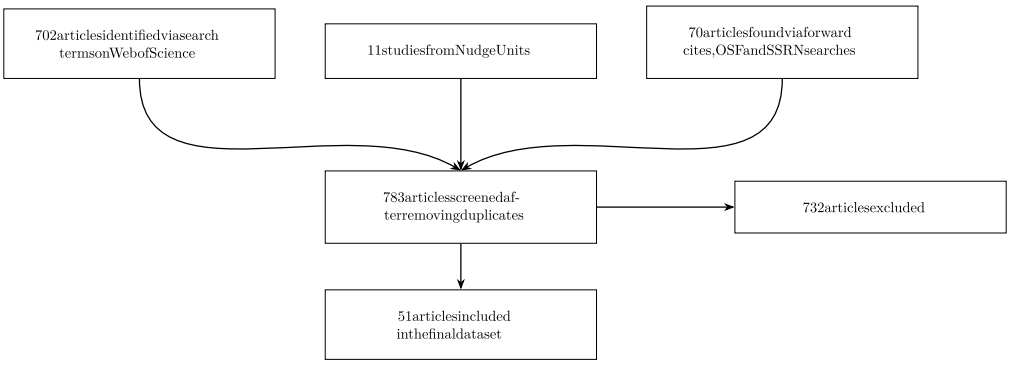


Figure 54: PRISMA-figure

Note: This figure records the original search. The dissertation update adds 15 studies and 46 effect sizes from the later adjudicated search.

Table 42: Included studies and effect sizes

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Milap and Sarin, 2023	Milap and Serin.2	Actual take-up	India	Free schooling	0.58	0.03	0.53	0.64	0.30
Milap and Sarin, 2023	Milap and Serin.1	Actual take-up	India	Free schooling	0.48	0.03	0.43	0.54	0.30
Fotuhi and Cohen, 2016	ideas_fotuhi.4	Actual take-up	USA (California)	Student benefit	0.42	0.10	0.22	0.62	0.10
Matikka and Paukkeri, 2022	Matikka, Tuomas and Paukkeri, Tuuli	Actual take-up	Finland	Pensioner benefit	0.35	0.00	0.35	0.36	0.10
Mueller and Yannelis, 2022	Mueller, Holger and Yannelis, Constantine	Actual take-up	USA (Nationwide)	Loan-forgiveness	0.34	0.01	0.32	0.36	0.10
Milap and Sarin, 2023	Milap and Serin.3	Actual take-up	India	Free schooling	0.32	0.03	0.26	0.38	0.30
Baillon et al., 2022	Baillon, Aurélien and Capuno, Joseph and O'Donnell, Owen and Tan, Carlos Antonio and van Wilgenburg, Kim	Actual take-up	Philippine	Subsidized health insurance	0.29	0.03	0.23	0.35	0.10
Fotuhi and Cohen, 2016	ideas_fotuhi.2	Actual take-up	USA (California)	Student benefit	0.29	0.11	0.07	0.51	0.00
Fotuhi and Cohen, 2016	ideas_fotuhi.3	Application	USA (California)	Student benefit	0.27	0.09	0.10	0.44	0.10
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.3	Application	USA (Ohio; North Carolina)	Student benefit	0.27	0.01	0.25	0.28	0.60
Asuming et al., 2024	Asuming, Patrick Opoku and Kim, Hyuncheol Bryant and Sim, Armand.3	Actual take-up	Ghana	Subsidized health insurance	0.23	0.03	0.17	0.29	0.30
Cid et al., 2019	Cid, Alejandro and Cabrera, José María and Bernatzky, Marianne and Ramírez-Michelena, María and Blanco, Magdalena.2	Actual take-up	Uruguay	Mix of multiple services	0.23	0.05	0.14	0.32	0.10
Van Gestel et al., 2023	Gestel, Raf Van and Goedemé, Tim and Janssens, Julie and Lefevere, Eva and Lemkens, Rik.2	Application	Belgium	Subsidized health insurance	0.23	0.01	0.22	0.24	0.60
Van Gestel et al., 2023	Gestel, Raf Van and Goedemé, Tim and Janssens, Julie and Lefevere, Eva and Lemkens, Rik.4	Application	Belgium	Subsidized health insurance	0.21	0.01	0.18	0.24	0.50
ideas42, 2015	ideas42_2015.3	Application	USA (Arizona)	Student benefit	0.21	0.00	0.20	0.22	0.50

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Milap and Sarin, 2023	Milap and Serin.4	Actual take-up	India	Free schooling	0.20	0.03	0.14	0.26	0.30
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.5	Application	USA (Ohio; North Carolina)	Student benefit	0.20	0.01	0.17	0.22	0.50
Schanzenbach, 2009	Diane Whitmore Schanzenbach.3	Application	USA (California)	Food stamps	0.18	0.04	0.11	0.25	0.20
Schanzenbach, 2009	Diane Whitmore Schanzenbach.4	Actual take-up	USA (California)	Food stamps	0.17	0.04	0.11	0.24	0.20
Van Gestel et al., 2023	Gestel, Raf Van and Goedemé, Tim and Janssens, Julie and Lefevere, Eva and Lemkens, Rik.6	Application	Belgium	Subsidized health insurance	0.17	0.00	0.16	0.18	0.60
Asuming et al., 2024	Asuming, Patrick Opoku and Kim, Hyuncheol Bryant and Sim, Armand.1	Actual take-up	Ghana	Subsidized health insurance	0.16	0.03	0.10	0.23	0.20
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.1	Application	USA (Ohio; North Carolina)	Student benefit	0.16	0.04	0.09	0.23	0.30
Van Gestel et al., 2023	Gestel, Raf Van and Goedemé, Tim and Janssens, Julie and Lefevere, Eva and Lemkens, Rik.1	Actual take-up	Belgium	Subsidized health insurance	0.15	0.01	0.14	0.16	0.60
Fotuhi and Cohen, 2016	ideas_fotuhi.1	Application	USA (California)	Student benefit	0.15	0.10	-0.05	0.35	0.10
ideas42, 2015	ideas42_2015.2	Application	USA (Arizona)	Student benefit	0.15	0.00	0.14	0.16	0.50
Wright et al., 2017	Wright, Bill J. and Garcia-Alexander, Ginny and Weller, Margarette A. and Baicker, Katherine.1	Actual take-up	USA (Oregon)	Subsidized health insurance	0.15	0.04	0.07	0.22	0.20
BETA, 2017	BETA.3	Application	Australia	Basic income	0.14	0.01	0.12	0.17	0.50
Wright et al., 2017	Wright, Bill J. and Garcia-Alexander, Ginny and Weller, Margarette A. and Baicker, Katherine.3	Actual take-up	USA (Oregon)	Subsidized health insurance	0.14	0.04	0.07	0.22	0.20
Cid et al., 2019	Cid, Alejandro and Cabrera, José María and Bernatzky, Marianne and Ramírez-Michelena, María and Blanco, Magdalena.1	Actual take-up	Uruguay	Mix of multiple services	0.14	0.04	0.06	0.22	0.10
BETA, 2017	BETA.1	Application	Australia	Basic income	0.14	0.01	0.12	0.17	0.50
Van Gestel et al., 2023	Gestel, Raf Van and Goedemé, Tim and Janssens, Julie and Lefevere, Eva and Lemkens, Rik.3	Actual take-up	Belgium	Subsidized health insurance	0.14	0.01	0.12	0.16	0.50

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Kilaru et al., 2026	Kilaru, Austin S. and Haider, Aliza and Harrison, Joseph and Dixon, Erica L. and Southwick, Lauren and Berkowitz, Melissa and Rareshide, Charles and Carroll, Conor and Kaledin, Clayton and McDermott, Grace and Mehta, Michael and Stephens Shields, Alisa J. and De La Rosa, Wendy and Agarwal, Anish K. and Merchant, Raina M.	Application	USA (Pennsylvania)	Mix of multiple services	0.14	0.04	0.06	0.21	0.10
BETA, 2017	BETA.7	Application	Australia	Basic income	0.13	0.01	0.11	0.16	0.50
BETA, 2017	BETA.9	Application	Australia	Basic income	0.13	0.01	0.11	0.16	0.50
BETA, 2017	BETA.5	Application	Australia	Basic income	0.13	0.01	0.10	0.16	0.50
BETA, 2017	BETA.11	Application	Australia	Basic income	0.13	0.01	0.10	0.16	0.50
Finkelstein and Notowidigdo, 2019	Finkelstein, Amy and Notowidigdo, Matthew J.1	Actual take-up	USA (Pennsylvania)	Food stamps	0.12	0.00	0.11	0.13	0.40
OES and USDT, 2024	OES4	Actual take-up	USA (Virginia)	Housing Benefit	0.12	0.00	0.11	0.13	0.10
ideas42, 2015	ideas42_2015.1	Application	USA (Arizona)	Student benefit	0.11	0.00	0.10	0.12	0.50
Engström et al., 2019	Engström, Per and Forsell, Eskil and Hagen, Johannes and Stefánsson, Arnaldur.4	Application	Sweden	Housing Benefit	0.11	0.01	0.09	0.12	0.60
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.2	Actual take-up	USA (Ohio; North Carolina)	Student benefit	0.11	0.03	0.04	0.17	0.30
Gupta, 2017	Gupta, Sarika.5	Application	India	Pensioner benefit	0.11	0.03	0.04	0.17	0.30
Van Gestel et al., 2023	Gestel, Raf Van and Goedemé, Tim and Janssens, Julie and Lefevere, Eva and Lemkens, Rik.5	Actual take-up	Belgium	Subsidized health insurance	0.10	0.00	0.09	0.11	0.60
Shen et al., 2024	Shen, Menghan and Tang, Zhiwei and Liang, Xiaoxia and Ye, Xiaoyang.2	Actual take-up	China	Subsidized health insurance	0.10	0.03	0.04	0.15	0.20
	Blanco, Mariana and Vargas, Juan F.	Actual take-up	Colombia	Mix of multiple services	0.09	0.06	-0.03	0.21	0.10
Engström et al., 2019	Engström, Per and Forsell, Eskil and Hagen, Johannes and Stefánsson, Arnaldur.8	Application	Sweden	Housing Benefit	0.09	0.01	0.08	0.10	0.60

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Avery et al., 2021	Christopher Avery Benjamin L. Castleman Michael Hurwitz Bridget T. Long Lindsay C. Page	Application	USA (Texas)	Student benefit	0.09	0.01	0.08	0.10	0.10
Hermes et al., 2025	Hermes, Henning and Lergetporer, Philipp and Peter, Frauke and Wiederhold, Simon.1	Application	Germany	Child Benefit	0.09	0.03	0.02	0.16	0.20
Engström et al., 2019	Engström, Per and Forsell, Eskil and Hagen, Johannes and Stefánsson, Arnaldur.6	Application	Sweden	Housing Benefit	0.09	0.01	0.08	0.10	0.60
Engström et al., 2019	Engström, Per and Forsell, Eskil and Hagen, Johannes and Stefánsson, Arnaldur.2	Application	Sweden	Housing Benefit	0.08	0.01	0.07	0.10	0.60
Fotuhi and Cohen, 2016	ideas_fotuhi.5	Application	USA (California)	Student benefit	0.08	0.08	-0.08	0.24	0.10
De La Rosa et al., 2021	De La Rosa, Wendy and Sharma, Eesha and Tully, Stephanie M. and Giannella, Eric and Rino, Gwen.2	Application	USA (California)	Tax credit	0.07	0.01	0.06	0.09	0.50
Gupta, 2017	Gupta, Sarika.6	Actual take-up	India	Pensioner benefit	0.07	0.03	0.01	0.13	0.30
Ravel et al., 2022	Ravel, Katie and Ahrary, Jahan and Avakian, Karen and Feher, Andrew and Menashe, Isaac.3	Actual take-up	USA (California)	Subsidized health insurance	0.07	0.03	0.02	0.12	0.30
Linos et al., 2024	Linos, Elizabeth and Reddy, Vikash and Rothstein, Jesse.1	Application	USA (California)	Student benefit	0.07	0.00	0.06	0.07	0.60
Hermes et al., 2025	Hermes, Henning and Lergetporer, Philipp and Peter, Frauke and Wiederhold, Simon.2	Actual take-up	Germany	Child Benefit	0.07	0.04	-0.02	0.15	0.20
Gupta, 2017	Gupta, Sarika.4	Actual take-up	India	Pensioner benefit	0.07	0.03	0.00	0.13	0.30
De La Rosa et al., 2021	De La Rosa, Wendy and Sharma, Eesha and Tully, Stephanie M. and Giannella, Eric and Rino, Gwen.1	Application	USA (California)	Tax credit	0.06	0.01	0.05	0.07	0.50
Moore et al., 2022	Moore, Ryan T. and Gan, Katherine N. and Minnich, Karissa and Yokum, David.2	Actual take-up	USA (Washington, DC)	Basic income	0.06	0.02	0.02	0.10	0.30
Kim et al., 2025	Kim, Jae Yeon and Herd, Pamela and Jilke, Sebastian and Moynihan, Donald and Rodden, Kerry	Actual take-up	USA (Colorado)	Food stamps	0.06	0.02	0.01	0.11	0.10
Giannella et al., 2023	Giannella, Eric and Homonoff, Tatiana and Rino, Gwen and Somerville, Jason	Actual take-up	USA (California)	Food stamps	0.06	0.00	0.05	0.07	0.10
Gupta, 2017	Gupta, Sarika.3	Application	India	Pensioner benefit	0.06	0.03	-0.00	0.12	0.30
Page et al., 2020	Page, Lindsay C. and Castleman, Benjamin L. and Meyer, Katharine	Application	USA (Texas)	Student benefit	0.06	0.01	0.04	0.07	0.10

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Cruces et al., 2020	Cruces, Guillermo and Amarante, Veronica and Ramírez, Lucía and Baez, Maria Josefina.2	Actual take-up	Argentina	Child Benefit	0.06	0.00	0.05	0.06	0.50
Linos et al., 2024	Linos, Elizabeth and Reddy, Vikash and Rothstein, Jesse.3	Application	USA (California)	Student benefit	0.06	0.00	0.05	0.06	0.60
Cruces et al., 2020	Cruces, Guillermo and Amarante, Veronica and Ramírez, Lucía and Baez, Maria Josefina.1	Actual take-up	Argentina	Child Benefit	0.05	0.00	0.05	0.06	0.50
Hemmeter et al., 2024	Jeffrey Hemmeter, John Phillips, Elana Safran, and Nicholas Wilson1*.5	Application	USA (Nationwide)	Pensioner benefit	0.05	0.00	0.05	0.05	0.60
OES and HHS, 2024	OES2.1	Actual take-up	USA (Minnesota)	Child Benefit	0.05	0.02	0.01	0.09	0.30
Hemmeter et al., 2024	Jeffrey Hemmeter, John Phillips, Elana Safran, and Nicholas Wilson1*.3	Application	USA (Nationwide)	Pensioner benefit	0.05	0.00	0.05	0.05	0.60
Engström et al., 2019	Engström, Per and Forsell, Eskil and Hagen, Johannes and Stefánsson, Arnaldur.3	Actual take-up	Sweden	Housing Benefit	0.05	0.00	0.04	0.06	0.60
Finkelstein and Notowidigdo, 2019	Finkelstein, Amy and Notowidigdo, Matthew J.2	Actual take-up	USA (Pennsylvania)	Food stamps	0.05	0.00	0.04	0.06	0.40
Gupta, 2017	Gupta, Sarika.2	Actual take-up	India	Pensioner benefit	0.05	0.03	-0.01	0.11	0.30
Cruces et al., 2020	Cruces, Guillermo and Amarante, Veronica and Ramírez, Lucía and Baez, Maria Josefina.3	Actual take-up	Argentina	Child Benefit	0.05	0.00	0.04	0.05	0.50
Hemmeter et al., 2024	Jeffrey Hemmeter, John Phillips, Elana Safran, and Nicholas Wilson1*.7	Application	USA (Nationwide)	Pensioner benefit	0.05	0.00	0.05	0.05	0.60
Moore et al., 2022	Moore, Ryan T. and Gan, Katherine N. and Minnich, Karissa and Yokum, David.1	Actual take-up	USA (Washington, DC)	Basic income	0.05	0.02	0.01	0.09	0.30
Guyton et al., 2017	Guyton, John and Langetieg, Pat and Manoli, Day and Payne, Mark and Schafer, Brenda and Sebastiani, Michael.2	Application	USA (Nationwide)	Tax credit	0.04	0.00	0.04	0.05	0.40
Hemmeter et al., 2024	Jeffrey Hemmeter, John Phillips, Elana Safran, and Nicholas Wilson1*.1	Application	USA (Nationwide)	Pensioner benefit	0.04	0.00	0.04	0.04	0.60

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Ravel et al., 2022	Ravel, Katie and Ahrary, Jahan and Avakian, Karen and Feher, Andrew and Menashe, Isaac.2	Actual take-up	USA (California)	Subsidized health insurance	0.04	0.02	-0.00	0.09	0.30
Engström et al., 2019	Engström, Per and Forsell, Eskil and Hagen, Johannes and Stefánsson, Arnaldur.7	Actual take-up	Sweden	Housing Benefit	0.04	0.00	0.03	0.05	0.60
Engström et al., 2019	Engström, Per and Forsell, Eskil and Hagen, Johannes and Stefánsson, Arnaldur.5	Actual take-up	Sweden	Housing Benefit	0.04	0.00	0.03	0.05	0.60
Ericson et al., 2023	Ericson, Keith Marzilli and Layton, Timothy J. and McIntyre, Adrianna and Sacarny, Adam.3	Actual take-up	USA (Massachusetts)	Subsidized health insurance	0.04	0.00	0.03	0.05	0.50
Carbuccia et al., 2025	Carbuccia, Laudine and Heim, Arthur and Barone, Carlo and Chevallier, Coralie.3	Application	France	Child Benefit	0.04	0.03	-0.01	0.09	0.30
Gupta, 2017	Gupta, Sarika.1	Application	India	Pensioner benefit	0.04	0.03	-0.02	0.10	0.30
Engström et al., 2019	Engström, Per and Forsell, Eskil and Hagen, Johannes and Stefánsson, Arnaldur.1	Actual take-up	Sweden	Housing Benefit	0.04	0.00	0.03	0.05	0.60
OES and HHS, 2024	OES2.2	Actual take-up	USA (Minnesota)	Child Benefit	0.04	0.02	-0.00	0.08	0.30
Banerjee et al., 2021	Banerjee, Abhijit and Finkelstein, Amy and Hanna, Rema and Olken, Benjamin A. and Ornaghi, Arianna and Sumarto, Sudarno.1	Actual take-up	Indonesia	Subsidized health insurance	0.04	0.02	0.00	0.07	0.40
Beecroft, 2012	Beecroft.2	Application	USA (Virginia)	Tax credit	0.03	0.01	0.01	0.06	0.50
Ericson et al., 2023	Ericson, Keith Marzilli and Layton, Timothy J. and McIntyre, Adrianna and Sacarny, Adam.2	Actual take-up	USA (Massachusetts)	Subsidized health insurance	0.03	0.00	0.02	0.04	0.50
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.4	Actual take-up	USA (Ohio; North Carolina)	Student benefit	0.03	0.01	0.02	0.04	0.60
Fotuhi and Cohen, 2016	ideas_fotuhi.6	Actual take-up	USA (California)	Student benefit	0.03	0.11	-0.19	0.25	0.00
Linos et al., 2024	Linos, Elizabeth and Reddy, Vikash and Rothstein, Jesse.5	Application	USA (California)	Student benefit	0.03	0.00	0.02	0.04	0.60
Lopoo et al., 2020	Lopoo, Leonard M. and Heflin, Colleen and Boskovski, Joseph.2	Actual take-up	USA (Minnesota)	Food stamps	0.03	0.01	0.01	0.05	0.40
Lopoo et al., 2020	Lopoo, Leonard M. and Heflin, Colleen and Boskovski, Joseph.3	Actual take-up	USA (Minnesota)	Food stamps	0.03	0.01	0.01	0.05	0.40

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Bai et al., 2021	Bai, Chong-En and Chi, Wei and Liu, Tracy Xiao and Tang, Chao and Xu, Jian.1	Actual take-up	China	Pensioner benefit	0.03	0.02	-0.01	0.07	0.40
Banerjee et al., 2021	Banerjee, Abhijit and Finkelstein, Amy and Hanna, Rema and Olken, Benjamin A. and Ornaghi, Arianna and Sumarto, Sudarno.3	Actual take-up	Indonesia	Subsidized health insurance	0.03	0.02	-0.01	0.07	0.40
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.11	Application	USA (Ohio; North Carolina)	Student benefit	0.03	0.02	-0.02	0.07	0.40
Myerson et al., 2022	Myerson, Rebecca and Tilipman, Nicholas and Feher, Andrew and Li, Honglin and Yin, Wesley and Menashe, Isaac	Actual take-up	USA (California)	Subsidized health insurance	0.03	0.00	0.02	0.03	0.10
Page et al., 2023	Page, Lindsay C. and Sacerdote, Bruce I. and Goldrick-Rab, Sara and Castleman, Benjamin L..3	Application	USA (Nationwide)	Student benefit	0.03	0.02	-0.01	0.06	0.40
Schanzenbach, 2009	Diane Whitmore Schanzenbach.1	Application	USA (California)	Food stamps	0.02	0.03	-0.04	0.09	0.30
Schanzenbach, 2009	Diane Whitmore Schanzenbach.2	Actual take-up	USA (California)	Food stamps	0.02	0.03	-0.04	0.09	0.30
Wei et al., 2026	Wei, Wenchi and Chen, Xin and Ni, Xing.3	Actual take-up	China	Tax credit	0.02	0.01	0.00	0.04	0.50
Lasky-Fink, 2023	Lasky-Fink, Jessica.4	Application	USA (California)	Food stamps	0.02	0.00	0.02	0.03	0.60
Lasky-Fink, 2023	Lasky-Fink, Jessica.6	Application	USA (California)	Food stamps	0.02	0.00	0.02	0.03	0.60
De La Rosa et al., 2021	De La Rosa, Wendy and Sharma, Eesha and Tully, Stephanie M. and Giannella, Eric and Rino, Gwen.4	Application	USA (California)	Tax credit	0.02	0.01	0.01	0.04	0.50
Wei et al., 2026	Wei, Wenchi and Chen, Xin and Ni, Xing.2	Actual take-up	China	Tax credit	0.02	0.01	0.00	0.04	0.50
DellaVigna and Linos, 2022	Nudge Unit	Actual take-up	USA	Tax credit	0.02	0.01	-0.01	0.05	0.10
Wei et al., 2026	Wei, Wenchi and Chen, Xin and Ni, Xing.1	Actual take-up	China	Tax credit	0.02	0.01	0.00	0.04	0.50
Lasky-Fink, 2023	Lasky-Fink, Jessica.2	Application	USA (California)	Food stamps	0.02	0.00	0.02	0.02	0.60
Bai et al., 2021	Bai, Chong-En and Chi, Wei and Liu, Tracy Xiao and Tang, Chao and Xu, Jian.2	Actual take-up	China	Pensioner benefit	0.02	0.02	-0.03	0.07	0.40

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Bai et al., 2021	Bai, Chong-En and Chi, Wei and Liu, Tracy Xiao and Tang, Chao and Xu, Jian.3	Actual take-up	China	Pensioner benefit	0.02	0.01	-0.00	0.04	0.50
Carbuccia et al., 2025	Carbuccia, Laudine and Heim, Arthur and Barone, Carlo and Chevallier, Coralie.4	Actual take-up	France	Child Benefit	0.02	0.03	-0.04	0.08	0.30
Wright et al., 2017	Wright, Bill J. and Garcia-Alexander, Ginny and Weller, Margarette A. and Baicker, Katherine.2	Actual take-up	USA (Oregon)	Subsidized health insurance	0.02	0.01	0.01	0.03	0.30
Hemmeter et al., 2024	Jeffrey Hemmeter, John Phillips, Elana Safran, and Nicholas Wilson1*.6	Actual take-up	USA (Nationwide)	Pensioner benefit	0.02	0.00	0.02	0.02	0.60
BETA, 2017	BETA.2	Actual take-up	Australia	Basic income	0.02	0.00	0.01	0.03	0.60
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.6	Actual take-up	USA (Ohio; North Carolina)	Student benefit	0.02	0.01	-0.00	0.04	0.60
Hemmeter et al., 2024	Jeffrey Hemmeter, John Phillips, Elana Safran, and Nicholas Wilson1*.4	Actual take-up	USA (Nationwide)	Pensioner benefit	0.02	0.00	0.02	0.02	0.60
Castell et al., 2022	Laura Castell Marc Gurgand Clement Imbert Todor Tochev.1	Actual take-up	France	Mix of multiple services	0.02	0.00	0.01	0.02	0.40
Linos et al., 2024	Linos, Elizabeth and Reddy, Vikash and Rothstein, Jesse.9	Application	USA (California)	Student benefit	0.02	0.00	0.01	0.02	0.60
OES and USDT, 2022	OES1.1	Actual take-up	USA (Nationwide)	Tax credit	0.02	0.01	-0.00	0.03	0.30
Wei et al., 2026	Wei, Wenchi and Chen, Xin and Ni, Xing.5	Actual take-up	China	Tax credit	0.02	0.01	-0.00	0.04	0.50
Hemmeter et al., 2024	Jeffrey Hemmeter, John Phillips, Elana Safran, and Nicholas Wilson1*.2	Actual take-up	USA (Nationwide)	Pensioner benefit	0.02	0.00	0.01	0.02	0.60
McQuillan and Moore, 2025	McQuillan, Casey and Moore, Brendan.3	Application	USA (Washington)	Basic income	0.02	0.00	0.01	0.02	0.50
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.12	Actual take-up	USA (Ohio; North Carolina)	Student benefit	0.01	0.02	-0.02	0.05	0.50
De La Rosa et al., 2021	De La Rosa, Wendy and Sharma, Eesha and Tully, Stephanie M. and Giannella, Eric and Rino, Gwen.3	Application	USA (California)	Tax credit	0.01	0.00	0.01	0.02	0.50
Domurat et al., 2021	Domurat, Richard and Menashe, Isaac and Yin, Wesley.4	Actual take-up	USA (California)	Subsidized health insurance	0.01	0.00	0.01	0.02	0.50

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Linos et al., 2025	LINOSLASKY2025.2	Application	USA (California)	Tax credit	0.01	0.00	0.01	0.02	0.60
Hemmeter et al., 2024	Jeffrey Hemmeter, John Phillips, Elana Safran, and Nicholas Wilson1*.8	Actual take-up	USA (Nationwide)	Pensioner benefit	0.01	0.00	0.01	0.01	0.60
OES and USDT, 2022	OES1.2	Actual take-up	USA (Nationwide)	Tax credit	0.01	0.01	-0.00	0.03	0.30
Shen et al., 2024	Shen, Menghan and Tang, Zhiwei and Liang, Xiaoxia and Ye, Xiaoyang.1	Actual take-up	China	Subsidized health insurance	0.01	0.02	-0.04	0.06	0.20
Feher et al., 2023	Feher, Andrew and Menashe, Isaac and Miller, Jennifer and Wolf, Emory.2	Actual take-up	USA (California)	Subsidized health insurance	0.01	0.00	0.00	0.02	0.40
BETA, 2017	BETA.4	Actual take-up	Australia	Basic income	0.01	0.00	0.00	0.02	0.60
Domurat et al., 2021	Domurat, Richard and Menashe, Isaac and Yin, Wesley.1	Actual take-up	USA (California)	Subsidized health insurance	0.01	0.00	0.01	0.02	0.50
Domurat et al., 2021	Domurat, Richard and Menashe, Isaac and Yin, Wesley.2	Actual take-up	USA (California)	Subsidized health insurance	0.01	0.00	0.01	0.02	0.50
Ericson et al., 2023	Ericson, Keith Marzilli and Layton, Timothy J. and McIntyre, Adrianna and Sacarny, Adam.1	Actual take-up	USA (Massachusetts)	Subsidized health insurance	0.01	0.00	0.00	0.02	0.50
Guyton et al., 2017	Guyton, John and Langetieg, Pat and Manoli, Day and Payne, Mark and Schafer, Brenda and Sebastiani, Michael.1	Application	USA (Nationwide)	Tax credit	0.01	0.00	0.01	0.02	0.40
Linos et al., 2025	LINOSLASKY2025.1	Application	USA (California)	Tax credit	0.01	0.00	0.01	0.01	0.60
Page et al., 2023	Page, Lindsay C. and Sacerdote, Bruce I. and Goldrick-Rab, Sara and Castleman, Benjamin L..1	Application	USA (Nationwide)	Student benefit	0.01	0.01	-0.01	0.04	0.40
BETA, 2017	BETA.12	Actual take-up	Australia	Basic income	0.01	0.00	0.00	0.02	0.60
Lopoo et al., 2020	Lopoo, Leonard M. and Heflin, Colleen and Boskovski, Joseph.1	Actual take-up	USA (Minnesota)	Food stamps	0.01	0.01	-0.00	0.03	0.40
McQuillan and Moore, 2025	McQuillan, Casey and Moore, Brendan.1	Application	USA (Washington)	Basic income	0.01	0.00	0.01	0.02	0.50
Banerjee et al., 2021	Banerjee, Abhijit and Finkelstein, Amy and Hanna, Rema and Olken, Benjamin A. and Ornaghi, Arianna and Sumarto, Sudarno.4	Actual take-up	Indonesia	Subsidized health insurance	0.01	0.01	-0.01	0.04	0.40
Linos et al., 2024	Linos, Elizabeth and Reddy, Vikash and Rothstein, Jesse.7	Application	USA (California)	Student benefit	0.01	0.00	0.00	0.02	0.60

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
OES et al., 2024	OES3.1	Application	USA (Texas)	Food stamps	0.01	0.00	0.00	0.02	0.40
Feher et al., 2023	Feher, Andrew and Menashe, Isaac and Miller, Jennifer and Wolf, Emory.1	Actual take-up	USA (California)	Subsidized health insurance	0.01	0.00	0.01	0.01	0.40
Domurat et al., 2021	Domurat, Richard and Menashe, Isaac and Yin, Wesley.3	Actual take-up	USA (California)	Subsidized health insurance	0.01	0.00	0.00	0.02	0.50
ideas42 and NYC Behavioral Design Team, 2022	ideas42_2022	Application	USA (New York)	Food stamps	0.01	0.00	0.00	0.02	0.10
Myerson et al., 2025	Myerson, Rebecca Mary and Espeseth, Allison and Dague, Laura	Actual take-up	USA (Wisconsin)	Subsidized health insurance	0.01	0.00	0.00	0.02	0.10
Beecroft, 2012	Beecroft.3	Application	USA (Virginia)	Tax credit	0.01	0.01	-0.01	0.03	0.50
Jones, 2010	Jones, Damon	Actual take-up	USA (Nationwide)	Tax credit	0.01	0.00	0.00	0.01	0.10
Linos et al., 2024	Linos, Elizabeth and Reddy, Vikash and Rothstein, Jesse.6	Actual take-up	USA (California)	Student benefit	0.01	0.00	0.00	0.02	0.60
BETA, 2017	BETA.8	Actual take-up	Australia	Basic income	0.01	0.00	-0.00	0.02	0.60
Goldin et al., 2021	Goldin, Jacob and Lurie, Ithai Z. and McCubbin, Janet	Actual take-up	USA (Nationwide)	Subsidized health insurance	0.01	0.00	0.01	0.01	0.10
Goldin et al., 2022	Goldin, Jacob and Homonoff, Tatiana and Javaid, Rizwan and Schafer, Brenda.1	Application	USA (Nationwide)	Tax credit	0.01	0.00	0.00	0.01	0.50
Asuming et al., 2024	Asuming, Patrick Opoku and Kim, Hyuncheol Bryant and Sim, Armand.2	Actual take-up	Ghana	Subsidized health insurance	0.01	0.02	-0.03	0.05	0.30
Beecroft, 2012	Beecroft.5	Application	USA (Virginia)	Tax credit	0.01	0.01	-0.01	0.02	0.50
BETA, 2017	BETA.10	Actual take-up	Australia	Basic income	0.01	0.00	-0.00	0.02	0.60
Linos et al., 2024	Linos, Elizabeth and Reddy, Vikash and Rothstein, Jesse.4	Actual take-up	USA (California)	Student benefit	0.01	0.00	-0.00	0.01	0.60
Beecroft, 2012	Beecroft.4	Application	USA (Virginia)	Tax credit	0.00	0.01	-0.01	0.02	0.50
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.4	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.00	0.01	0.70
Linos et al., 2025	LINOSLASKY2025.5	Application	USA (California)	Tax credit	0.00	0.00	0.00	0.01	0.60
OES et al., 2024	OES3.2	Actual take-up	USA (Texas)	Food stamps	0.00	0.00	-0.01	0.02	0.40

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
McQuillan and Moore, 2025	McQuillan, Casey and Moore, Brendan.4	Actual take-up	USA (Washington)	Basic income	0.00	0.00	0.00	0.01	0.50
Yokum et al., 2022	Yokum, David and Hopkins, Daniel J. and Feher, Andrew and Safran, Elana and Peck, Joshua.2	Actual take-up	USA (Nationwide (37 HealthCare.gov states))	Subsidized health insurance	0.00	0.00	0.00	0.01	0.60
Yokum et al., 2022	Yokum, David and Hopkins, Daniel J. and Feher, Andrew and Safran, Elana and Peck, Joshua.4	Actual take-up	USA (Nationwide (37 HealthCare.gov states))	Subsidized health insurance	0.00	0.00	0.00	0.01	0.60
Amirapu et al., 2022	Amirapu, Amrit and Clots-Figueras, Irma and Malde, Bansi and Mitra, Anirban and Pakrashi, Debayan and Wahhaj, Zaki.2	Actual take-up	India	Food stamps	0.00	0.02	-0.03	0.04	0.20
Banerjee et al., 2021	Banerjee, Abhijit and Finkelstein, Amy and Hanna, Rema and Olken, Benjamin A. and Ornaghi, Arianna and Sumarto, Sudarno.2	Actual take-up	Indonesia	Subsidized health insurance	0.00	0.01	-0.02	0.03	0.40
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.10	Actual take-up	USA (Ohio; North Carolina)	Student benefit	0.00	0.01	-0.02	0.03	0.50
Goldin et al., 2022	Goldin, Jacob and Homonoff, Tatiana and Javaid, Rizwan and Schafer, Brenda.2	Actual take-up	USA (Nationwide)	Tax credit	0.00	0.00	0.00	0.01	0.50
Goldin et al., 2022	Goldin, Jacob and Homonoff, Tatiana and Javaid, Rizwan and Schafer, Brenda.3	Actual take-up	USA (Nationwide)	Tax credit	0.00	0.00	0.00	0.01	0.50
Goldin et al., 2022	Goldin, Jacob and Homonoff, Tatiana and Javaid, Rizwan and Schafer, Brenda.4	Actual take-up	USA (Nationwide)	Tax credit	0.00	0.00	0.00	0.00	0.50
Linos et al., 2024	Linos, Elizabeth and Reddy, Vikash and Rothstein, Jesse.2	Actual take-up	USA (California)	Student benefit	0.00	0.00	-0.00	0.01	0.60
Linos et al., 2025	LINOSLASKY2025.6	Application	USA (California)	Tax credit	0.00	0.00	-0.00	0.01	0.60
Page et al., 2023	Page, Lindsay C. and Sacerdote, Bruce I. and Goldrick-Rab, Sara and Castleman, Benjamin L..2	Application	USA (Nationwide)	Student benefit	0.00	0.02	-0.03	0.04	0.40
McQuillan and Moore, 2025	McQuillan, Casey and Moore, Brendan.2	Actual take-up	USA (Washington)	Basic income	0.00	0.00	0.00	0.01	0.50
Yokum et al., 2022	Yokum, David and Hopkins, Daniel J. and Feher, Andrew and Safran, Elana and Peck, Joshua.3	Actual take-up	USA (Nationwide (37 HealthCare.gov states))	Subsidized health insurance	0.00	0.00	0.00	0.01	0.60

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Lasky-Fink, 2023	Lasky-Fink, Jessica.5	Application	USA (California)	Food stamps	0.00	0.03	-0.05	0.06	0.40
Castell et al., 2022	Laura Castell Marc Gurgand Clement Imbert Todor Tochev.2	Actual take-up	France	Mix of multiple services	0.00	0.00	-0.01	0.01	0.40
Linos et al., 2024	Linos, Elizabeth and Reddy, Vikash and Rothstein, Jesse.10	Actual take-up	USA (California)	Student benefit	0.00	0.00	-0.00	0.01	0.60
Yokum et al., 2022	Yokum, David and Hopkins, Daniel J. and Feher, Andrew and Safran, Elana and Peck, Joshua.7	Actual take-up	USA (Nationwide (37 HealthCare.gov states))	Subsidized health insurance	0.00	0.00	0.00	0.00	0.60
Lasky-Fink, 2023	Lasky-Fink, Jessica.1	Application	USA (California)	Food stamps	0.00	0.03	-0.05	0.06	0.40
Lasky-Fink, 2023	Lasky-Fink, Jessica.3	Application	USA (California)	Food stamps	0.00	0.03	-0.05	0.06	0.40
Lasky-Fink and Linos, 2024	Lasky-Fink, Jessica and Linos, Elizabeth.1	Application	USA (Colorado)	Housing Benefit	0.00	0.00	0.00	0.00	0.40
Page et al., 2023	Page, Lindsay C. and Sacerdote, Bruce I. and Goldrick-Rab, Sara and Castleman, Benjamin L..4	Application	USA (Nationwide)	Student benefit	0.00	0.02	-0.03	0.03	0.40
Yokum et al., 2022	Yokum, David and Hopkins, Daniel J. and Feher, Andrew and Safran, Elana and Peck, Joshua.1	Actual take-up	USA (Nationwide (37 HealthCare.gov states))	Subsidized health insurance	0.00	0.00	0.00	0.00	0.60
Yokum et al., 2022	Yokum, David and Hopkins, Daniel J. and Feher, Andrew and Safran, Elana and Peck, Joshua.5	Actual take-up	USA (Nationwide (37 HealthCare.gov states))	Subsidized health insurance	0.00	0.00	0.00	0.00	0.60
Lasky-Fink and Linos, 2024	Lasky-Fink, Jessica and Linos, Elizabeth.2	Application	USA (Colorado)	Housing Benefit	0.00	0.00	-0.00	0.00	0.40
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.3	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.00	0.01	0.70
Linos et al., 2025	LINOSLASKY2025.3	Application	USA (California)	Tax credit	0.00	0.00	0.00	0.00	0.60
Linos et al., 2025	LINOSLASKY2025.4	Application	USA (California)	Tax credit	0.00	0.00	0.00	0.00	0.60
Yokum et al., 2022	Yokum, David and Hopkins, Daniel J. and Feher, Andrew and Safran, Elana and Peck, Joshua.6	Actual take-up	USA (Nationwide (37 HealthCare.gov states))	Subsidized health insurance	0.00	0.00	-0.00	0.00	0.60
Yokum et al., 2022	Yokum, David and Hopkins, Daniel J. and Feher, Andrew and Safran, Elana and Peck, Joshua.8	Actual take-up	USA (Nationwide (37 HealthCare.gov states))	Subsidized health insurance	0.00	0.00	-0.00	0.00	0.60

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
BETA, 2017	BETA.6	Actual take-up	Australia	Basic income	0.00	0.01	-0.01	0.01	0.60
Hervelin, 2022	Hervelin, Jérémy.1	Actual take-up	France	Mix of multiple services	-0.00	0.02	-0.04	0.04	0.30
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.2	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.00	0.00	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.5	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.01	0.01	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.6	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.01	0.01	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.8	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.00	0.00	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.9	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.00	0.00	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.12	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.00	0.00	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.15	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.00	0.00	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.16	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.00	0.00	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.18	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.00	0.00	0.70

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.19	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.01	0.01	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.20	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.01	0.01	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.21	Actual take-up	USA (California)	Tax credit	0.00	0.00	-0.01	0.01	0.70
Bai et al., 2021	Bai, Chong-En and Chi, Wei and Liu, Tracy Xiao and Tang, Chao and Xu, Jian.4	Actual take-up	China	Pensioner benefit	0.00	0.02	-0.03	0.03	0.40
Nguyen et al., 2014	Wagstaff, Adam and Nguyen, Ha Thi Hong and Dao, Huyen and Bales, Sarah	Actual take-up	Vietnam	Subsidized health insurance	0.00	0.01	-0.02	0.02	0.10
Bergman et al., 2019	Bergman, Peter and Denning, Jeffrey T. and Manoli, Dayanand.1	Application	USA (Texas)	Student benefit	-0.00	0.00	-0.01	0.01	0.50
Bergman et al., 2019	Bergman, Peter and Denning, Jeffrey T. and Manoli, Dayanand.2	Application	USA (Texas)	Student benefit	-0.00	0.00	-0.00	0.00	0.50
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.1	Actual take-up	USA (California)	Tax credit	-0.00	0.00	-0.00	0.00	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.10	Actual take-up	USA (California)	Tax credit	-0.00	0.00	-0.00	0.00	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.11	Actual take-up	USA (California)	Tax credit	-0.00	0.00	-0.00	0.00	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.13	Actual take-up	USA (California)	Tax credit	-0.00	0.00	-0.00	0.00	0.70
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.14	Actual take-up	USA (California)	Tax credit	-0.00	0.00	-0.00	0.00	0.70

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.17	Actual take-up	USA (California)	Tax credit	-0.00	0.00	-0.00	0.00	0.70
Bergman et al., 2019	Bergman, Peter and Denning, Jeffrey T. and Manoli, Dayanand.3	Application	USA (Texas)	Student benefit	-0.00	0.00	-0.00	0.00	0.50
Linos et al., 2022	Linos, Elizabeth and Prohofsky, Allen and Ramesh, Aparna and Rothstein, Jesse and Unrath, Matthew.7	Actual take-up	USA (California)	Tax credit	-0.00	0.00	-0.00	-0.00	0.70
Rogers, 2025	Rogers, Parker	Actual take-up	USA (California)	Food stamps	-0.00	0.01	-0.01	0.01	0.10
Wei et al., 2026	Wei, Wenchi and Chen, Xin and Ni, Xing.4	Actual take-up	China	Tax credit	-0.00	0.01	-0.02	0.02	0.50
Amirapu et al., 2022	Amirapu, Amrit and Clots-Figueras, Irma and Malde, Bansi and Mitra, Anirban and Pakrashi, Debayan and Wahhaj, Zaki.1	Actual take-up	India	Food stamps	-0.00	0.03	-0.07	0.06	0.20
Ravel et al., 2022	Ravel, Katie and Ahrary, Jahan and Avakian, Karen and Feher, Andrew and Menashe, Isaac.1	Actual take-up	USA (California)	Subsidized health insurance	-0.00	0.02	-0.04	0.04	0.30
Bird et al., 2021	Bird, Kelli A. and Castleman, Benjamin L. and Denning, Jeffrey T. and Goodman, Joshua and Lamberton, Cait and Rosinger, Kelly Ochs	Application	USA (Texas)	Student benefit	-0.01	0.00	-0.01	-0.00	0.10
Beecroft, 2012	Beecroft.1	Application	USA (Virginia)	Tax credit	-0.01	0.01	-0.04	0.02	0.50
Linos et al., 2024	Linos, Elizabeth and Reddy, Vikash and Rothstein, Jesse.8	Actual take-up	USA (California)	Student benefit	-0.01	0.00	-0.01	0.00	0.60
Bai et al., 2021	Bai, Chong-En and Chi, Wei and Liu, Tracy Xiao and Tang, Chao and Xu, Jian.5	Actual take-up	China	Pensioner benefit	-0.01	0.02	-0.04	0.02	0.40
Carbuccia et al., 2025	Carbuccia, Laudine and Heim, Arthur and Barone, Carlo and Chevallier, Coralie.1	Application	France	Child Benefit	-0.01	0.03	-0.06	0.04	0.30
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.7	Application	USA (Ohio; North Carolina)	Student benefit	-0.01	0.06	-0.13	0.10	0.20
Hervelin, 2022	Hervelin, Jérémy.2	Actual take-up	France	Mix of multiple services	-0.01	0.01	-0.04	0.02	0.30

Continued on next page

Paper 4: Included studies and effect sizes (continued)

Study	Study arm	Outcome	Country	Benefit type	Effect	SE	CI low	CI high	Weight
De La Rosa et al., 2021	De La Rosa, Wendy and Sharma, Eesha and Tully, Stephanie M. and Giannella, Eric and Rino, Gwen.5	Application	USA (California)	Tax credit	-0.01	0.01	-0.03	-0.00	0.50
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.8	Actual take-up	USA (Ohio; North Carolina)	Student benefit	-0.02	0.06	-0.12	0.09	0.20
Bettinger et al., 2012	Bettinger, Eric P. and Long, Bridget Terry and Oreopoulos, Philip and Sanbonmatsu, Lisa.9	Application	USA (Ohio; North Carolina)	Student benefit	-0.02	0.01	-0.05	0.01	0.50
Carbuccia et al., 2025	Carbuccia, Laudine and Heim, Arthur and Barone, Carlo and Chevallier, Coralie.2	Actual take-up	France	Child Benefit	-0.02	0.03	-0.08	0.04	0.30
Pepin et al., 2021	G Pepin, C O'Leary, D Oberlee	Application	USA (Michigan)	Basic income	-0.03	0.04	-0.10	0.04	0.10

G.2 Further descriptives

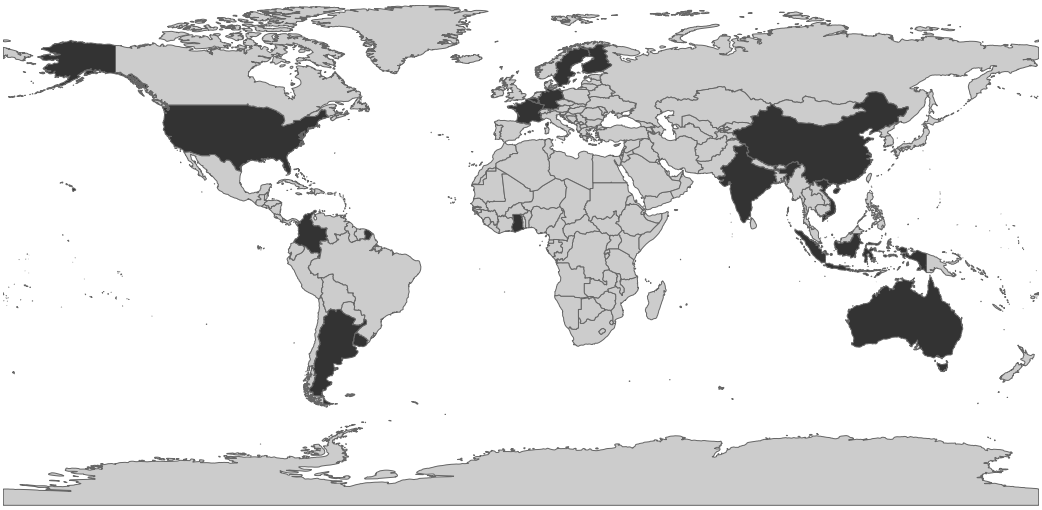


Figure 55: Map of the included countries

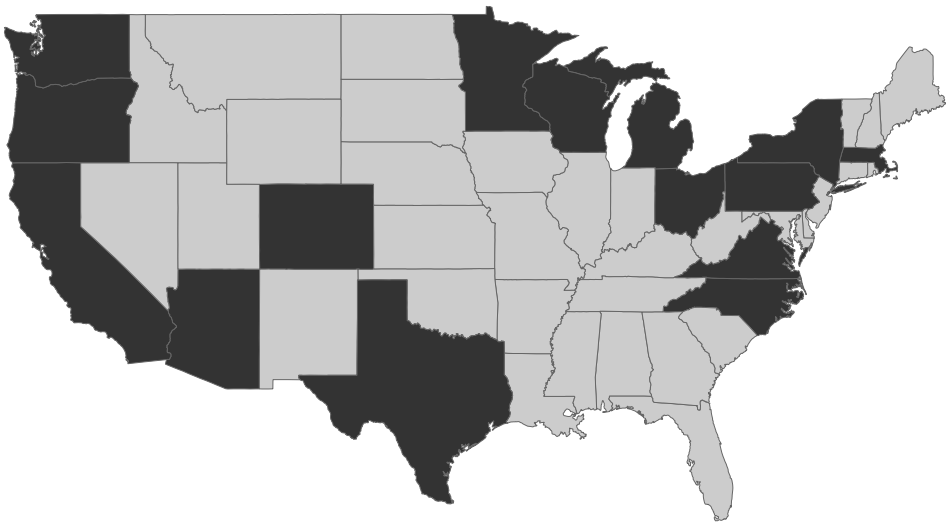


Figure 56: Map of included states

State within	Effect sizes	Studies
California	70	13
Nationwide	23	8
North Carolina	12	1
Ohio	12	1
Nationwide (37 HealthCare.gov states)	8	1
Texas	8	5
Virginia	6	2
Minnesota	5	2
Washington	4	1
Arizona	3	1
Colorado	3	2
Massachusetts	3	1
Oregon	3	1
Pennsylvania	3	2
Washington, DC	2	1
Michigan	1	1
New York	1	1
Wisconsin	1	1

G.3 Search string

TO = ((RCT OR "field experiment*" OR "randomized controlled") AND (take-up OR uptake) AND (social benefits OR social program OR government OR Social Security OR health insurance OR SNAP OR TANF OR EITC OR MEDICAID OR MEDICARE OR HOUSING BENEFIT* OR "HOUSING ALLOWANCE*" OR "Housing Assistance" OR "voucher" OR "food stamp*" OR "Social Assistance" OR "In-work Benefit*" OR "Income Support" OR "Unemployment Benefit*" OR "Family Benefit*" OR "Minimum Pension Supplement" OR "Income Supplement" OR "Means-tested Benefit*" OR "Tax Benefit" OR "Tax credit" OR "Child Benefit" OR "Pension Credit" OR "Child Support Grant" OR "Cash Transfer*"))

G.4 Publication bias plots and tables

Below is the funnel plot for all effect sizes.

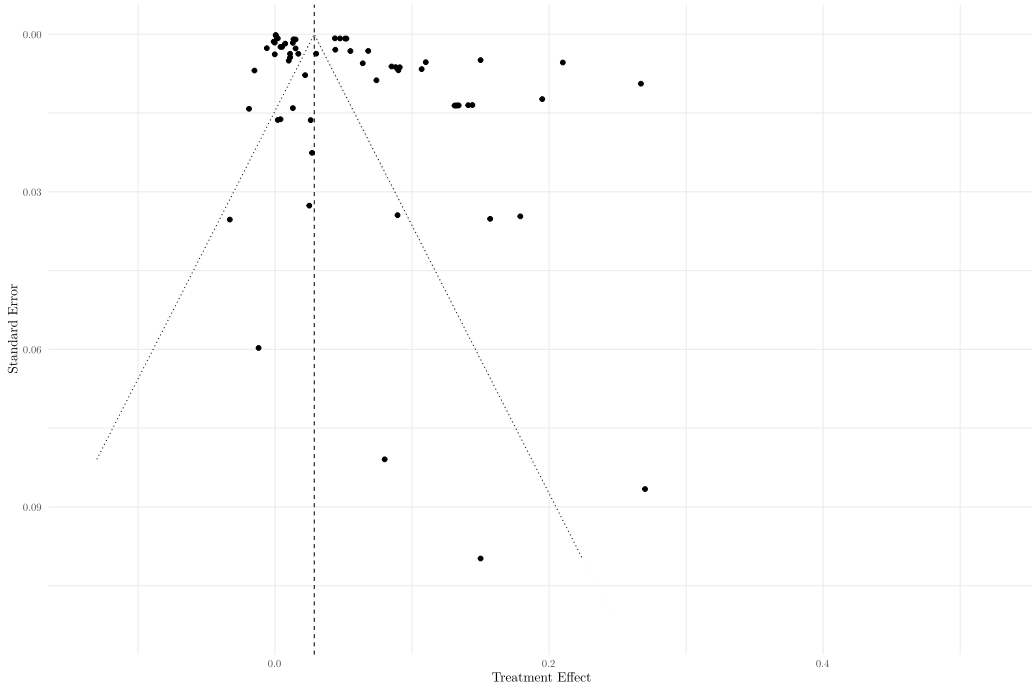


Figure 57: Funnel-plot of effect sizes

Andrews & Kasy models

Below are the models based on Andrews and Kasy (2019): first for all effect sizes and second for all effect sizes measuring actual take-up.

Θ	Σ	β_p
0.021	0.045	0.121
(0.0086)	(0.0138)	(0.0687)

Across all models

Θ	Σ	β_p
0.016	0.039	0.083
(0.0064)	(0.0124)	(0.0358)

Only measuring actual take-up

Θ	Σ	β_p
-9.0e-01	3.3e-05	5.2e-01
(0.29)	(4345.81)	(0.37)

Only measuring application *Note:* The application-only optimization is numerically unstable in the updated sample, as indicated by the implausible corrected mean and extremely large uncertainty estimate. It should not be interpreted substantively.

G.5 Regression tables

Overall average effect

	(1)
overall	0.060***
	(0.011)
Num.Obs.	233

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 43: Meta-regression output of all studies. SEs are CR2

Risk ratio

	RR model	RR interaction model
Intercept (Actual take-up, Learning)	0.244***	0.382***
	(0.045)	(0.069)
Application	0.144**	0.141
	(0.044)	(0.141)
Num.Obs.	232	232
SE type	CR2	CR2
Cluster level	Study	Study

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 44: Meta-regression using log(RR) as outcome

Note: The outcome-only model estimates risk ratios of 1.28 for actual take-up and 1.47 for application. In the interaction model, the estimated risk ratios are 1.22 for learning-demand interventions measuring actual take-up, 1.47 for compliance-demand interventions measuring actual take-up, 1.41 for learning-demand interventions measuring application, and 1.69 for compliance-demand interventions measuring application.

G.6 Interaction with variable for extra steps

For every application effect size, I looked at whether the people had to complete further steps in the application process or whether only the eligibility check was left. While this analysis is very underpowered, the direction of the estimate suggests that there are larger effect sizes in studies that measure earlier in the application process, such as on initiation or application to SNAP or TANF before the interviews. There is still, however, a positive effect of measuring application, even when there are no extra steps.

	(1)
intercept	0.044***
	(0.012)
ApplicationvstakeupApplication	0.035
	(0.016)
extrasteps	0.017
	(0.025)
Num.Obs.	233

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 45: Meta-regression of different application steps

Note: Studies reducing compliance demands are excluded

Differences in learning demands interventions?

There are many different ways that researchers have tried to increase take-up using simple nudges. As these are rarely the same, it is difficult to test many of the differences. However, there are many studies claiming to reduce the psychological costs of applying for a benefit. Testing the differences within the studies that try to reduce learning demands suggests that there are no discernible differences between the different types of "nudges". This might be because the idea of reducing psychological costs is very broad and involves both reducing stigma and giving psychological ownership. It could also suggest that the level of psychological costs experienced varies considerably between programs, so only some of the treatments will have an effect.

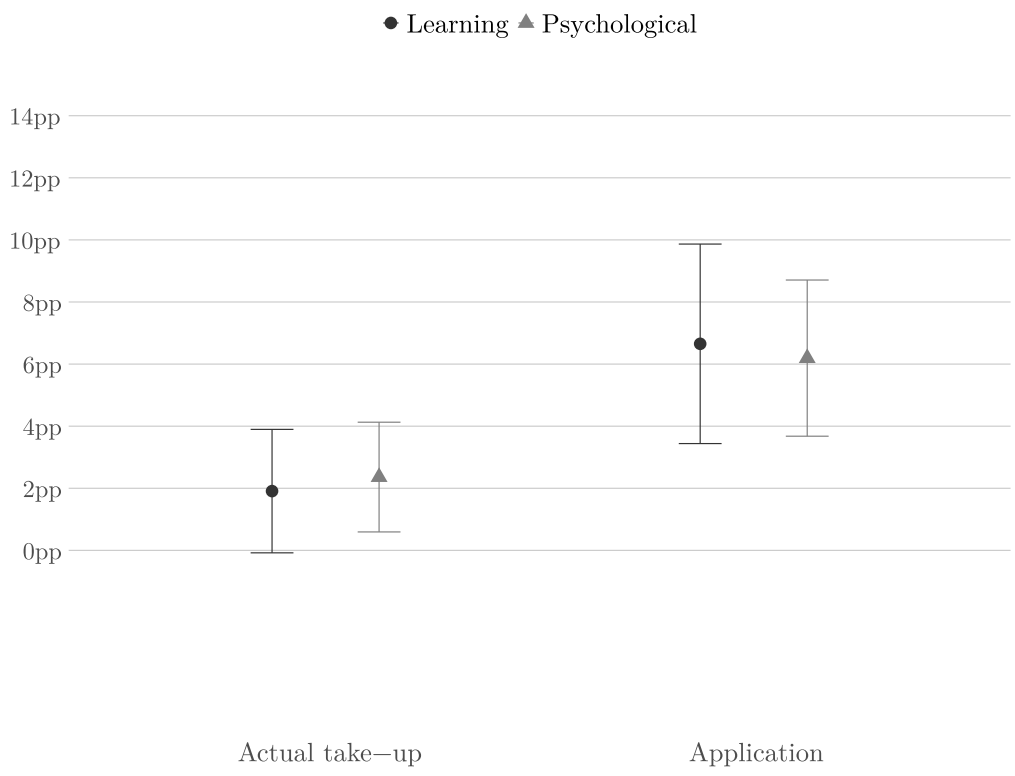

Learning
Psychological
14pp
12pp
10pp
8pp
6pp
4pp
2pp
0pp
Actual take−up
Application

Figure 58

Difference across control base-level

One might initially expect that studies with low baseline take-up rates in the control group would show larger treatment effects, as there is more room for improvement and a relatively larger pool of individuals who can respond to the intervention. However, the analysis reveals no systematic difference in effect sizes based on the initial take-up rates in the control group. The effect sizes are the largest in the studies where the control group take-up rate is around 20–40%. This replicates the findings from DellaVigna and Linos (2022). Looking at the risk ratio plot, it is clear that effects with a very low take-up rate in the control group are highly inflated compared to the other studies. There are even a few studies with risk ratios of over 5. This effect size would be impossible for 54% of the studies with a control group take-up rate of over 20 percent, as there is a natural ceiling at 100% take-up.

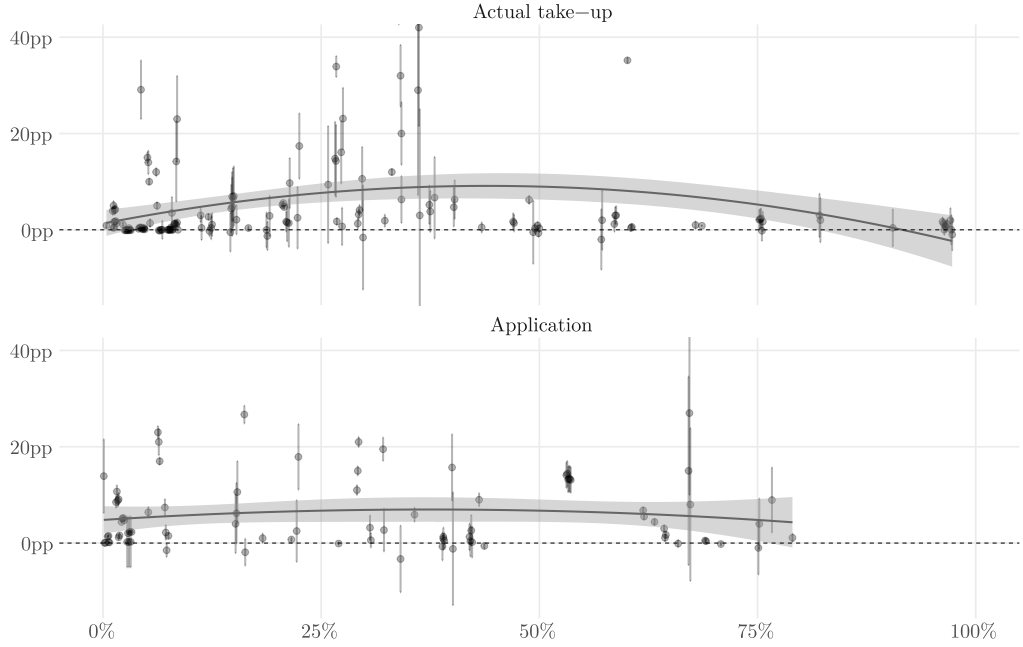


Figure 59

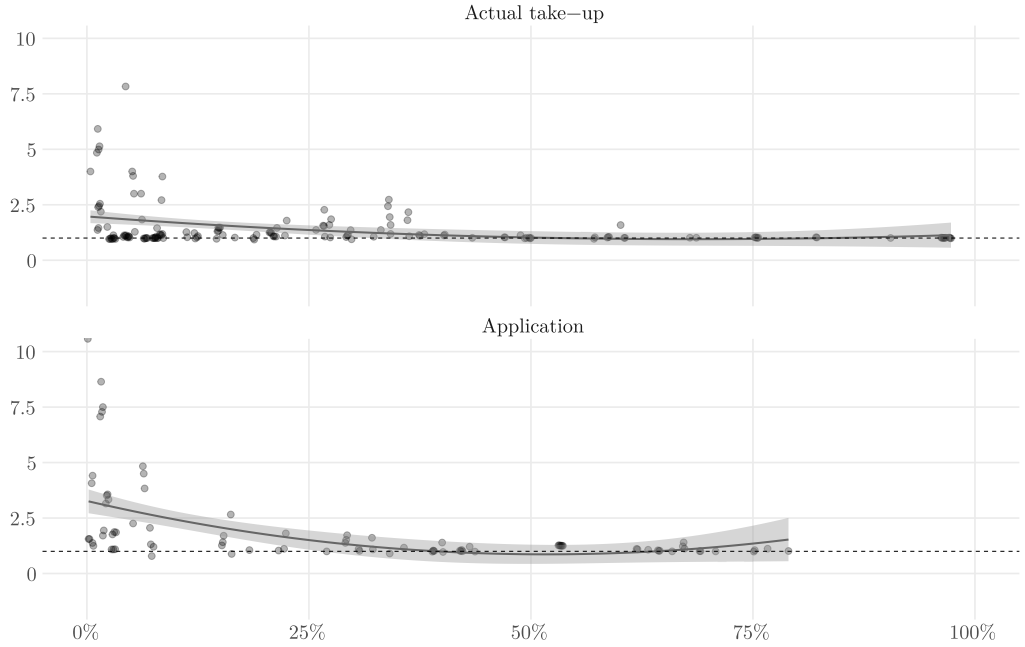


Figure 60

G.7 Further notes on estimation

I use a three-level random-effects model to synthesize results from the field experiments aimed at increasing the take-up of social benefits. This approach is common within meta-analyses (Harrer et al., 2021), and I use it to account for the hierarchical structure of the data, where effect sizes are nested within studies, and studies are nested within broader contexts, such as institutional settings and types of benefits. Thus, I estimate a model with three different error terms for the estimated true effect across settings.

At the first level of the model, the observed effect size $\hat{\theta}_{ij}$ for each intervention i within a study j is modeled as a function of the true effect size θ_{ij} and a random error term ϵ_{ij}. This level is expressed as:

$$\hat{\theta}_{ij} = \theta_{ij} + \epsilon_{ij} \tag{28}$$

where ϵ_{ij} represents the within-study sampling error.

At the second level, the true effect size θ_{ij} varies across different interventions within the same study. The variability in effect sizes within each study is captured by the model:

$$\theta_{ij} = \kappa_j + \zeta_{ij}^{(2)} \tag{29}$$

where κ_j represents the average effect size within study j, and $\zeta_{ij}^{(2)}$ is the deviation of intervention i from this study average.

At the third level, I account for the variability in average effect sizes across different studies. The study-level effect size κ_j is modeled as:

$$\kappa_j = \mu + \zeta_j^{(3)} \tag{30}$$

where μ represents the overall average effect size across all studies, and $\zeta_j^{(3)}$ captures the deviation of study j from this overall average.

By combining these three levels, the full model can be expressed as:

$$\hat{\theta}_{ij} = \mu + \zeta_j^{(3)} + \zeta_{ij}^{(2)} + \epsilon_{ij} \tag{31}$$

This model incorporates two heterogeneity components: $\zeta_{ij}^{(2)}$, representing the variation in true effect sizes within studies, and $\zeta_j^{(3)}$, representing the variation between studies. The three-level meta-analysis model was fitted using restricted maximum likelihood (REML) estimation, which was implemented in the *metafor* package within R. Using this type of model, I account for differences in power within the studies so that studies with higher numbers of observations contribute most to estimates. If not specified, I used approximations of the number of observations given the specified percentages selected for each group.

As some studies find effects and others do not, this model simply averages effect sizes across studies without explaining the heterogeneity in the findings. It is therefore more interesting for this paper to estimate differences between studies by including an interaction term for various study characteristics X_{ij}, such as the outcome measured.

$$\hat{\theta}_{ij} = \mu + \beta X_{ij} + \zeta_j^{(3)} + \zeta_{ij}^{(2)} + \epsilon_{ij} \tag{32}$$