Regulatory Compliance with Limited Enforceability: Evidence from Privacy Policies*

Bernhard Ganglmair† Julia Krämer‡ Jacopo Gambato§

October 12, 2023

Abstract

The EU General Data Protection Regulation (GDPR) of 2018 introduced stringent transparency rules compelling firms to disclose the nature of their data collection, processing, and use in accessible and readable language. The disclosure requirement is objective, and its compliance is verifiable. However, readability is subjective and vague, making it difficult to enforce. We examine the effect of this asymmetric enforceability of regulatory rules on firms' compliance using a large sample of privacy policies from German firms between 2014 and 2021, matched with firm-level and industry-level information. We use text-as-data techniques to construct measures of disclosure and readability and show that firms responded to the GDPR's transparency requirements by significantly increasing information disclosure. However, the readability of their privacy policies did not improve and, in some cases, worsened. Larger firms and those in concentrated industries demonstrated higher compliance levels with the readability requirement, potentially due to heightened regulatory scrutiny. We emphasize the significance of regulatory capacity, as higher-budget regulators (German state-level data protection authorities) with better enforcement capabilities foster improved compliance with the vague rules and guidance of the readability requirement. This study sheds light on the intricate dynamics between enforceability, compliance, and the role of verifiability within regulatory frameworks.

Keywords: data protection, disclosure, GDPR, privacy policies, readability, regulation, text-as-data, topic models

JEL Codes: C81; D23; K12; K20; L51; M15.

*The order of authors is randomized using the AEA's Author Randomization Tool. We thank Kirsten Bock, Jaroslaw Kantorowicz, Wolfgang Kerber, Jan Krämer, and Alexander Wulf and conference and seminar participants at the University of Antwerpen, the European Association of Law and Economics, the Society of Institutional and Organizational Economics, and the German Law and Economics Association for useful comments and suggestions. Many thanks to Ryan Amos for sharing the data of the Princeton-Leuven Longitudinal Corpus of Privacy Policies. We also thank Jianming Cui, Natalia Garcia Soto, Pujit Golchha, Ana Rantes Lozano, and Lion Szlagowski for excellent research assistance. Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (Project B2) is gratefully acknowledged. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

†University of Mannheim and ZEW Mannheim, b.ganglmair@gmail.com

‡Erasmus University Rotterdam, j.k.kramer@law.eur.nl

§University of Mannheim and ZEW Mannheim, jacopo.gambato@zew.de

1 Introduction

In 2018, the General Data Protection Regulation went into effect; it transformed the digital landscape in Europe and beyond, often to the detriment of firms but with some privacy improvement on the consumer side.[1] A central contribution of the GDPR was its transparency requirement that compels firms to disclose information about the nature of their data collection, processing, and use (Art. 13–14 GDPR) in accessible and readable language (Art. 12(1) GDPR). However well-intended, the rules come with major enforceability concerns. While the disclosure requirement is based on an objective list of items to be disclosed, the readability requirement is vague and subjective; data protection authorities are left to interpret these rules as they lack enforcement experience and established precedent or best practices.[2] Because enforcement is costly, data protection authorities will prioritize and eventually give more weight to the disclosure requirement in their enforcement activities.[3] But if firms anticipate limited enforcement in one dimension, compliance will suffer. This paper asks how the asymmetric (and limited) enforceability of the GDPR's transparency requirement affects firms' compliance decisions. We also explore in greater detail the role of regulatory scrutiny (when firms anticipate they are a regulator's primary target) and regulatory capacity (Stern, 2000; Armstrong and Sappington, 2006) in the strategic interaction of enforcement and compliance.

We first propose a theoretical model to address these questions. In our framework, a firm can choose costly compliance with some requirements when drafting a privacy policy, and a regulator can audit the privacy policy to confirm compliance. Our framework is closest to that of Heyes (1994), which models the thoroughness of inspection of a single requirement as an endogenous choice. We deviate from this approach by taking the probabilities of success (with less than certainty) of an audit in detecting non-compliance with multiple requirements as given (Macho-Stadler and Perez-Castrillo, 2006), and instead focus on the firm's and regulator's choices of which of the requirements to comply with and audit. We assume asymmetric enforcement success with a higher probability of detecting non-compliance for the disclosure requirement and a lower probability for the readability requirement. We derive equilibrium outcomes for constrained regulators (that can enforce only one of the requirements) and unconstrained regulators (that can enforce one or both requirements).

This model of asymmetric enforcement imperfections predicts better disclosure com-

[1] See Johnson (forthcoming) for a comprehensive review of the economics literature studying the effect of the GDPR

[2] Art. 29 Working Party (2018, para. 9) (the predecessor of the European Data Protection Board) refers to "readability tests" as a means of enforcement, without providing specific details (and further guidance).

[3] DataGrail (2019, 5) paints a colorful picture: "Benchmarking the financial cost of compliance as a baseline, 74% of companies spent more than $100,000 on compliance consulting services and technology solutions, and 20% spent more than $1M. One-third (34%) of enterprise companies (1,000+ employees) spent more than $1M.

pliance than readability compliance as a firm's response to the GDPR (Prediction 1). Moreover, when a firm expects to be a more prominent target for the regulator and thus anticipates stricter enforcement (*regulatory scrutiny*), it will showcase better readability compliance than disclosure compliance (Prediction 2). Intuitively, such a firm already exhibits robust disclosure compliance, and higher regulatory scrutiny triggers a catching-up effect (i.e., better readability compliance). Last, relaxing the regulator's budget constraint (and allowing for broader enforcement through increased *regulatory capacity*) again triggers a catching-up effect. A firm facing an unconstrained regulator will exhibit better readability compliance than a firm subject to a constrained regulator (Prediction 3). A firm with high compliance costs will reduce its disclosure compliance in response.

To test the predictions from our model, we construct a quarterly (unbalanced) panel of privacy policies posted by German firms between 2014 and 2021. We restrict the sample to firms with at least one privacy policy before the GDPR became enforceable (Q2 2018) and one policy after. The final dataset contains more than 585,000 policies posted by more than 75,000 firms. For Prediction 1, we conduct a simple before-and-after analysis. For our empirical test of Prediction 2, we use firm size and market concentration as proxies for regulatory scrutiny. We match observations in the policy panel with information on firm size (number of employees), market concentration (4-digit industry Herfindahl-Hirschman Index), and firm address that we obtain from the Mannheim Enterprise Panel (Bersch et al., 2014). For Prediction 3, we exploit that Germany has 16 state data protection authorities that regulate the firms in their respective states. We exploit the variation of the authorities' budget both across states and over time, assuming that higher-budget data protection authorities are less likely to be budget constrained.[4] To this end, we collect budget information for the 16 German state data protection authorities (2012–2022) and use state-level population numbers from the German Federal Statistical Office to calculate per-capita budget variables.

For our outcome variables of a firm's compliance with the transparency requirements, we construct metrics for disclosure and readability. We use standard methods from natural language processing to construct measures for simple informational volume and length of privacy policies (e.g., number of words and sentences). We estimate LDA topic models (Blei et al., 2003) and count a policy's distinct topics to capture the breadth of policies. We finally use these topic model results to identify paragraphs using terms indicative of the disclosure of information required by Art. 13–14 GDPR (i.e., disclosing paragraphs). To construct our primary disclosure measure (topic-weighted word count), we use the relative distribution of topics of these disclosing paragraphs as weights of a

[4]Enforcement is costly for regulators, and data protection authorities vary in resources – both in Germany and across Europe (European Data Protection Board, 2020), they also vary in strictness and tactics. For instance, survey evidence suggests the strictest regulators are found in Germany and Sweden (see Johnson, forthcoming).

policy's paragraphs' word count.

For readability measurements, we borrow from the toolkit of linguists, who have constructed many indices and scores to measure the readability of texts. We use two scores. First, as best-practice approach, we construct the German version of the Flesch Reading Ease (German FRE) score (Flesch, 1948; Amstad, 1978) that has been used in the U.S. to regulate the readability of insurance contracts. Second, we take a data-driven approach following Benoit et al. (2019). We compile a set of roughly 4,000 comparisons (by humans) of the readability of short snippets of text from our sample of privacy policies. Using the methodology laid out in Benoit et al. (2019), we then identify the *läsbarhetsindex* (LIW) (Björnson, 1968) as the readability index that best explains our comparison data.

We find that, in response to the GDPR, privacy policies increase in length by 50–70%. The amount of disclosure (in disclosing paragraphs) increases by almost 80%. We, therefore, find strong evidence for disclosure compliance, whereas the results for readability compliance are weak and mixed. The GDPR response for the German FRE score indicates a decline in readability and for the LIW an increase in readability. Both are an order of magnitude (measured in percent or impact on a human's ability to differentiate two texts by their readability) smaller than the effects on disclosure. These findings are in line with Prediction 1 of our theoretical framework. They also highlight the "inherent tension" within the transparency requirement to provide more information in a more concise form under the GDPR (Art. 29 Working Party, 2018, para 34). We further explore the impact of the GDPR on disclosure and readability as a function of a firm's exposure to the GDPR (or the respective intensity). We find that firms with pre-GDPR policies that disclosed little or showed little readability (and were, therefore, more exposed to the requirements of the GDPR) exhibited more robust increases in disclosure and readability (for both measures). Put differently, firms with relatively readable privacy policies paid less attention to the readability of their post-GDPR privacy policies than firms that had to catch up.

We find empirical support for Prediction 2 of our model for the German FRE but not the LIW index. Larger firms and those in higher-concentration markets show better readability compliance. In the case of the German FRE, readability declines less for firms subject to more stringent regulatory scrutiny. The absence of the result for the LIW might have to do with expectations. The FRE has a regulatory track record (in the context of insurance regulation in the U.S.). For lack of better guidance, if firms expect data protection authorities to enforce the readability requirement, they may also expect the use of established and tested measures. While the LIW performs better as a readability measure in our specific text domain, it does not have the same regulatory track record.

The results on disclosure compliance are mixed. Higher regulatory scrutiny by market

concentration does not affect a firm's disclosure response to the GDPR. Larger firms, however, respond with even more disclosure than small firms. An alternative mechanism (other than regulatory scrutiny) is that for larger firms, the drafting and compliance costs are relatively lower than for small firms. Also, larger firms (and their legal counsel) may be more likely to draft customized privacy policies that are expected to be longer than off-the-shelf templates.[5]

Our regulator budget results are partially in line with Prediction 3. We find that firms in higher-budget states do not exhibit different levels of disclosure compliance. We obtain noisy negative coefficients on the interaction terms. If anything, these firms' disclosure compliance declines (a result we obtain for firms with high compliance costs). The results on readability compliance are strongest for the German FRE. Firms in higher-budget states (total budget per capita and labor budget per capita) see a smaller decline in readability than firms with more constrained data protection authorities. We do not see the same patterns for the LIW.

The remainder of this paper is structured as follows. In Section 3, we provide background information on the transparency regime in the GDPR. In Section 4, we introduce our theoretical framework and derive predictions for the empirical analysis. In Section 5, we describe the construction of our estimation sample. In Section 6, we dive into the construction of text-based measures of disclosure and readability. In Section 7, we document how firms have responded to the introduction of the GDPR using simple before-and-after analyses and a treatment-intensity approach. In Section 8, we explore the role of regulatory scrutiny and capacity. Finally, we conclude in Section 9.

2 Related Literature

Our study contributes to a broad literature. A growing number of studies examine the effects of the GDPR on firm behavior and performance. Examples are Yuan and Li (2019) (a sharp decline in financial performance for hospitals that attach importance to digital health services), Goldberg et al. (2019) (a 13.3% drop of revenue for e-commerce sites), or Johnson et al. (forthcoming) and Peukert et al. (2022) (examining the effects of the GDPR on firms' use of and interaction with web technology vendors). Koski and Valmari (2020) find that small and medium-sized enterprises in data-intensive industries are affected the most by the GDPR, arguing that economies of scale may result in different economic effects of the GDPR when adhering to its regulations. Our results on the effects of firm size on compliance back these findings. We also add to this literature by providing a nuanced picture of the effectiveness of the GDPR and highlighting that compliance with

[5]The findings in Ganglmair and Wardlaw (2017) suggest that when contracting parties customize the text of loan agreements, the result tends to be shorter than standardized or boilerplate contract language.

the new regulation is not a given but rather the result of firms' anticipation of strategic enforcement decisions by constrained regulators.

The literature studying privacy policies has seen a sharp increase in attention with the introduction of the GDPR. While earlier work uses small samples of privacy policies (e.g., Jensen and Potts, 2004; Milne et al., 2006), more recent studies compile large datasets of privacy policies for thousands of firms. For instance, Frankenreiter (2022) uses a curated sample of 60,000 privacy policies (from 700 websites) in the U.S. (demonstrating that the GDPR had a limited effect on U.S. businesses, as most analyzed policies have not been updated accordingly), and Amos et al. (2021) or Wagner (2023) examine more than 1 million policies spanning a window of over two decades (examining longer-term changes in content and readability of policies). Both small and large-scale studies have found a downward trend in the readability of privacy policies (e.g., Milne et al., 2006; Amos et al., 2021), with some recent results also hinting at no-changes (Linden et al., 2020) or slight improvements (Becher and Benoliel, 2021) post-GDPR. Moreover, studies show that post-GDPR, privacy policies are significantly longer and show greater detail (Degeling et al., 2019; Linden et al., 2020). We can match privacy policies to firm and industry-level data and thus paint a more nuanced picture of trends in informational volume, disclosure, and readability. Moreover, our approach provides evidence for limited enforceability (and resulting lack of enforcement) as a potential explanation for the failure of the GDPR to provide more readable (and transparent) privacy policies.[6]

The nature and characteristics of regulatory environments have been the subject of a line of study in the literature on regulation economics. Systematic limitations to regulation generally relate to information asymmetries between the regulator and the regulated industry (Laffont, 1994) and limited regulatory resources (Stern, 2000; Armstrong and Sappington, 2006).[7] We focus our attention on a different kind of impediment to regulation, namely the limited enforceability of uncertain or vague requirements. Uncertainty of regulatory requirements can interact with a regulator's budget constraint, mainly when regulation is multi-dimensional (with several requirements that must be met), and limited resources reduce the ability of a regulator to enforce them all. Our empirical results show that the effect of the budget is disproportionate for the requirement that comes with a higher level of vagueness or lower verifiability.

Our theoretical framework builds on the game-theoretical literature on audits and tax avoidance (Greenberg, 1984; Fellingham and Newman, 1985; Graetz et al., 1986). We follow in their footsteps and include a novel dimension to the strategy of the regulator. Rather than focusing on the optimal intensity of the auditing efforts, we consider the optimal regulator strategy with regard to what she should audit when agents are compelled

[6]See the Commission's 2019 status report (European Commission, 2019).

[7]These limitations are typically (but not exclusively) studied in the context of developing countries where the premises for perfect enforcement are not generally met (Stern, 2000; Laffont, 2005).

to comply with multiple requirements. Unlike earlier work, we model imperfect regulation (Heyes, 1994; Bardsley, 1996; Macho-Stadler and Perez-Castrillo, 2006)[8] and assume exogenous success probabilities (as in Macho-Stadler and Perez-Castrillo, 2006). We add to this literature by studying the enforcement and compliance of multiple requirements that must be audited separately with different success probabilities.

Our results also relate to the literature on contractual terms of use, "fine print," and boilerplate (or standardized) contract language. Bakos et al. (2014) show overwhelming evidence supporting the notion that users rarely even skim through the fine print of contracts and terms of use. Given this lack of attention by consumers, Marotta-Wurgler (2008) asks if firms display more or less predatory contracts in response to the clients' disregard of the contracts. She shows a striking homogeneity in contract terms. Our findings differ in this respect. We show that privacy policies exhibit a high level of heterogeneity in both length and text structure. The findings in Ganglmair and Wardlaw (2017) (studying the texts of large bank loans) suggest that when contracting parties customize the text of agreements, the result tends to be shorter than standardized or boilerplate contract language. We find evidence in line with these findings.

Last, our methodological approach relates our study to a growing literature that uses text-as-data methods. For a comprehensive survey of this growing literature, see Loughran and McDonald (2016) (in finance and accounting) or Gentzkow et al. (2019) (in the social sciences). A central method in our paper is the estimation of topic models. These models have been used on a number of different types of document corpora such as emails (McCallum et al., 2007), scientific abstracts (Blei et al., 2003; Griffiths and Steyvers, 2004) and articles (Hall et al., 2008; Blei, 2012), newspaper archives (Wei and Croft, 2006; Larsen and Thorsrud, 2019), U.S. Supreme Court decisions (Livermore et al., 2016), patents (Ruckman and McCarthy, 2017), loan agreements (Ganglmair and Wardlaw, 2017), or analyst reports (Ball et al., 2015; Bellstam et al., 2021). We add privacy policies to this ever-growing list.

3 Transparency Requirements in the GDPR

The General Data Protection Regulation (GDPR) contains a set of cumulative principles that are a prerequisite for any form of processing of personal data and ensure their lawful processing. One of these principles is *transparency* (in Art. 5(1) lit. a GDPR) which requires any information concerning the processing of personal data to be easily accessible and understandable. The underlying aim behind this principle is that consumers need to understand the information provided to them to be able to make informed decisions about who and how their data are processed. For this reason, Art. 12 GDPR specifies the

[8] We also assume audits (or regulation inspection) are error-free, which means, they do not produce false negatives by mistaking compliance for non-compliance.

procedural and technical aspects of information provision: The form and communication of this information are crucial for consumers to be able to assess the consequences of the processing taking place (Polčák, 2020, 402). Art. 12 goes hand in hand with Art. 13 and 14 GDPR, which require a firm to provide consumers with details about the data processing: Consumers cannot effectively exercise their rights without complete information (Zanfir-Fortuna, 2020, 416).[9,10]

A legal obligation to provide transparent information to users is not an entirely new concept in the EU legal order. Information rights for data subjects have their origins in the German concept of informational self-determination, which the German Federal Constitutional Court established in the famous census ruling (Polčák, 2020, 401). The right to informational self-determination guarantees the power of the individual to determine the disclosure and use of his or her personal data.[11] The EU Data Protection Directive (DPD), which regulated the processing of personal data before the GDPR, already referred to concepts that relate to the regulation of information provision.[12] While Art. 12 DPD was primarily concerned with substantive rights, it provided that information had to be delivered in an "intelligible form."

The scope and reach of the GDPR, however, is broader than that of the DPD. As stated in Art. 12(1) GDPR, all information duties must meet the general standard of informing the data subject "in a concise, transparent, intelligible and easily accessible form, using clear and plain language." In addition to the introduction of a new transparency standard, Art. 13(1) GDPR sets out new information that has to be delivered to a data subject. These duties include, for example, informing users of the identity of the data controller (Art. 4(7) GDPR) or the choice of the legal basis underlying data collection before processing their personal data (Art. 6(1) lit. a–f GDPR). The DPD had similar provisions in Art. 10 DPD, but those were limited to information on data processing purposes and the data-collecting firm's identity.

The two faces of the transparency principle – *readability* in Art. 12 and *disclosure* in Art. 13 and 14 – may seem to point in the same direction when in fact, they are said to be in conflict with each other. The Art. 29 Working Party (2018, para. 34) observes "an inherent tension in the GDPR between the requirements on the one hand to provide

[9]Elements that have to be disclosed include, for instance, the contact details of a firm, the legal basis the data processing is based on, and the duration of the storage of personal data.

[10]In the United States, the California Consumer Privacy Act (CCPA) also requires companies to disclose information to users in privacy policies (Cal. Civ. Code §1798.130(a)(5)(A)-(C)). Among the information that must be disclosed are the purpose of the data collection and a description of consumers' rights. While these obligations overlap with the GDPR, the CCPA does not require firms to draft readable privacy policies. Only the information requested from users must be answered in a format "easily understandable by the average consumer" and, if possible, "machine-readable." Consequently, the transparency principle is broader in the EU than in the United States.

[11]Bundesverfassungsgericht (Federal Constitutional Court), December 15, 1983, 65, 1, 42.

[12]The DPD entered into force in 1995 and was implemented into German law via the *Telemediengesetz* (TMG), which codifies the duty to inform the data subject about the nature, scope, and purpose of the collection and use of personal data (§13(1) TMG).

the comprehensive information to data subjects which is required under the GDPR, and on the other hand do so in a format that is concise, transparent, intelligible and easily accessible." In addition to this apparent tradeoff that complicates compliance with the GDPR's transparency principle, the subjectivity of what it means for the form of information to be "concise, transparent, intelligible and easily accessible" creates another layer of enforcement issues.

While the information disclosure requirements in Art. 13 and 14 are straightforward (as the required content is spelled out in a detailed list that is objectively verifiable), the lack of explicitly stated and verifiable measures for Art. 12 raises compliance concerns. It should come as no surprise that transparency provisions in data protection law, and specifically in the GDPR, have a long history of criticism for being ineffective in enhancing the privacy offered to users (see, e.g., Solove (2013), Waldman (2021, 61 ff.)).

The Art. 29 Working Party (2018) has tried to address the ensuing enforcement and compliance issues by providing non-binding guidelines to facilitate a consistent application of the law. It gave terms such as "concise and transparent," "intelligible," and "clear and plain language" a more precise definition. It also emphasized the needs of the "average member of the intended audience" and how that average user ought to be able to easily access information expressed in "as simple a manner as possible." As a standard to assess the compliance with Art. 12, the Art. 29 Working Party (2018, para. 9) has proposed mechanisms such as "readability testing."[13] For that purpose, readability indices and scores are usually constructed using various linguistic components. The Working Party, however, does not provide guidance on which components should be considered relevant and why, and which index or score is the most suitable for the analysis of legal documents such as privacy policies. Also, because the European Court of Justice has not yet clarified how to assess compliance with Art. 12(1) GDPR, the Working Party's suggestions serve (at best) as loose guidance for firms and data protection authorities.

4 A Simple Model of Compliance

We propose a simple model to capture a firm's decision to comply with the requirements of the GDPR. In our framework, a firm can choose costly compliance with some requirements when drafting a privacy policy, and a regulator can audit the privacy policy to confirm compliance. Our framework is closest to that of Heyes (1994), which models thoroughness of inspection of a single requirement as an endogenous choice. We deviate from this approach by taking the probability of success of an audit in detecting non-compliance with

[13]Paragraph 9 reads: "If controllers are uncertain about the level of intelligibility and transparency of the information and effectiveness of user interfaces/ notices/ policies etc., they can test these, for example, through mechanisms such as user panels, readability testing, formal and informal interactions and dialogue with industry groups, consumer advocacy groups and regulatory bodies, where appropriate, amongst other things."

multiple requirements as given Macho-Stadler and Perez-Castrillo (2006), and instead focus on the choice of which of the requirements to comply with and to audit. We study the role that firm size and the expected level of received scrutiny play in how firms draft their policies. Furthermore, we highlight the role of a regulator's budget constraint. Through the model, we make several predictions that we bring to the data.

4.1 Framework

A firm (she) is tasked with drafting a privacy policy, subject to two requirements. First, it must provide the right type and amount of information ("disclosure"); second, the policy must be accessible to consumers ("readability"). Compliance with these requirements is costly for the firm. A regulator (he) is tasked with enforcing the disclosure and readability requirements. He audits policies to assess their compliance. The regulator can choose the intensity of this audit and inspect the policy for either, neither, or both requirements. Audits are imperfect: the regulator learns, with positive probability, whether the policy complies with either requirement. If he finds non-compliance, he challenges the policy, resulting in a penalty for the firm.

We model the interaction between the firm and the regulator as a simultaneous-move game in which the firm chooses how many and which requirements to comply with, and the regulator chooses the intensity of the audit (that is, which requirements, if any, to inspect).

The firm's stand-alone value from an unchallenged policy is $v > 0$. Compliance with each requirement $j \in \{d, r\}$ (for *disclosure* and *readability*) comes at a fixed cost k per requirement. Formally, the firm selects $(d, r) \in \{0, 1\} \times \{0, 1\}$ and generates an unchallenged value $v - kd - kr$. If the regulator audits the policy and finds non-compliance, the firm's payoffs are zero.[14] We assume that compliance never leads to negative payoffs for the firm:

Assumption 1. $0 < k < \frac{v}{2}$

For an audit, the regulator chooses to inspect either, neither, or both requirements. When he finds non-compliance in either requirement, he challenges the policy. Both audit and challenge are without cost. We consider two types of regulators: *unconstrained* and *constrained*. An unconstrained regulator has sufficient resources to inspect both requirements. A constrained regulator has limited resources and can inspect at most one

[14] A challenged policy always generates zero utility for the firm, regardless of its choice of (d, r). This implies that the fee paid by the firm for non-compliance is different if she does not comply with either or both requirements. The assumption allows for immediate comparison of all possible outcomes. A flat fee would not affect the compliance incentives beyond a numerical difference. We model the payoffs this way to highlight the interaction between the firm's and regulator's choices rather than produce direct numerical estimates.

requirement.[15]

Non-compliant policies generate a social loss of $-\gamma < 0$.[16] The regulator's objective is to minimize this social loss, and he audits policies to detect non-compliance, subject to its constraint type. His payoffs from an unchallenged non-compliant policy are $-\gamma$; the payoffs from a challenged or compliant policy are zero.

Audits are imperfect, and the inspection of a policy for requirement j leads to the discovery of its state (either $d, r = 1$ or $d, r = 0$) with probability π_j. We assume that inspecting disclosure d has a higher chance of discovering the true state of the policy. This is motivated by our earlier discussion: disclosure of specific information items is objective, whereas readability is subjective. We further introduce a lower bound for the regulator's success probabilities.[17]

Assumption 2. $\frac{1}{2} < \pi_r < \pi_d < 1$

The strategies of the players are as follows: The firm chooses between full non-compliance ($d = 0, r = 0$), non-compliance in readability ($d = 1, r = 0$), non-compliance in disclosure ($d = 0, r = 1$), and full compliance ($d = 1, r = 1$). To ease notation, we refer to the strategic decision $j = 1$ as j, and $j = 0$ as 0, for $j \in \{d, r\}$.

An unconstrained regulator chooses between no inspection (a_0), inspection of the disclosure requirement (a_d), inspection of the readability requirement (a_r), and full inspection (of both requirements) ($a_{d,r}$). A constrained regulator has only single-inspection choices and cannot choose full inspection. Table 1 summarizes the players' strategies and corresponding outcomes. The first value in each cell represents the firm's payoffs; the second value represents the regulator's payoffs.

4.2 Equilibrium

We first derive the Nash equilibrium for the unconstrained regulator (Proposition 1) and then proceed to the constrained regulator (Proposition 2). Last, we compare regulatory environments with unconstrained relative to constrained regulators (Proposition 3).

4.2.1 Unconstrained Regulator

In Table 1, we can see that if the regulator is unconstrained, he chooses to inspect both requirements as $a_{d,r}$ is a dominant strategy. Given this dominant choice by the regulator,

[15] This regulator-type distinction is a reduced-form characterization of a regulator's budget constraint. Suppose inspecting a given requirement comes at a cost, say, c. Then, an unconstrained regulator has sufficient resources to incur costs $2c$, whereas a constrained regulator can afford only audit costs of c.

[16] We assume a non-compliant policy generates the same social loss $-\gamma$ for any form of non-compliance. This assumption is for simplicity and not an assessment of the social loss from lack of disclosure relative to lack of readability.

[17] This second assumption ensures that, for all feasible values k (Assumption 1), the firm does not always strictly prefer not to comply with either requirement.

Table 1: Normal-Form Representation of the Compliance-Enforcement Game

		Regulator's strategy			
		a_0	a_d	a_r	$a_{d,r}$
Firm's strategy	$(0,0)$	$\begin{pmatrix} v \\ -\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_d)v \\ -(1-\pi_d)\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_r)v \\ -(1-\pi_r)\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_d)(1-\pi_r)v \\ -(1-\pi_d)(1-\pi_r)\gamma \end{pmatrix}$
	$(d,0)$	$\begin{pmatrix} v-k \\ -\gamma \end{pmatrix}$	$\begin{pmatrix} v-k \\ -\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_r)(v-k) \\ -(1-\pi_r)\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_r)(v-k) \\ -(1-\pi_r)\gamma \end{pmatrix}$
	$(0,r)$	$\begin{pmatrix} v-k \\ -\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_d)(v-k) \\ -(1-\pi_d)\gamma \end{pmatrix}$	$\begin{pmatrix} v-k \\ -\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_d)(v-k) \\ -(1-\pi_d)\gamma \end{pmatrix}$
	(d,r)	$\begin{pmatrix} v-2k \\ 0 \end{pmatrix}$	$\begin{pmatrix} v-2k \\ 0 \end{pmatrix}$	$\begin{pmatrix} v-2k \\ 0 \end{pmatrix}$	$\begin{pmatrix} v-2k \\ 0 \end{pmatrix}$

the firm always prefers $(d,0)$ to $(0,r)$. If she decides to comply with only one requirement, she optimally chooses to comply with what is easier to detect (e.g., disclosure). Moreover, by the lower bound of the regulator's success probability in Assumption 2, the firm either chooses full compliance (d,r) or non-compliance in readability $(d,0)$.

Proposition 1 (Unconstrained Regulator)**.** *Suppose the regulator is unconstrained and can inspect both requirements. Let* $k^u = \frac{\pi_r}{1+\pi_r}v$*. For low compliance costs with* $k < k^u$*, the equilibrium is* $(a_{d,r},(d,r))$*. For high compliance costs with* $k \geq k^u$*, the equilibrium is* $(a_{d,r},(d,0))$*. In both cases, the regulator inspects both requirements and the firm always complies with disclosure. The firm also complies with readability when enforcement costs are low.*

Proof. The proof is relegated to Appendix A. □

The extent to which the firm complies with *both* d and r depends on the compliance cost k. Because $\pi_d > \pi_r$, it is always better for the firm to comply with the disclosure requirement than with the readability requirement. It is also strictly better than not complying at all under the assumption that compliance never leads to negative payoffs for the firm. Lastly, if k is small enough, full compliance is cheap and the benefits (by avoiding a challenge by the regulator) outweigh the costs. Conversely, if compliance costs k are high, the firm prefers not to comply with r, hoping that her non-compliance goes undetected. The threshold value at which the firm is indifferent between these options, k^u, is increasing in π_r. All else equal, an increase in π_r implies that the firm strictly prefers to comply with both requirements for more values of k. In the limit, $\pi_r = 1$ leads to $k^u = \frac{v}{2}$, and the firm always complies with both requirements when audits are perfect (by Assumption 1 it holds $k < \frac{v}{2}$).

4.2.2 Constrained Regulator

For a constrained regulator, a full audit with $a_{d,r}$ is not feasible. However, given cost-free audits, an audit with *some* inspection dominates no audit. As a consequence, the regulator selects either a_d or a_r, possibly using a mixed strategy.

It is straightforward to see that a pure-strategy equilibrium does not exist. Suppose the regulator selects to inspect disclosure, a_d, with probability one. The firm's best response is to choose $(d, 0)$, that is, comply with respect to disclosure and ignore the readability requirement. The regulator is then unable to challenge the non-complying firm and would want to deviate, choosing a_r instead to be able to challenge the policy (that is not readability compliant). And so forth. In equilibrium, the regulator will always play a mixed strategy, choosing a_d and a_r with strictly positive probabilities.

The firm does not want to comply with both requirements if she can avoid it. To find the respective equilibria, we proceed as follows: First, we obtain the firm's mixed strategies with probabilities of playing $(d, 0)$ (denoted by p_d), $(0, r)$ (denoted by p_r), and $(0, 0)$ (probability $1 - p_d - p_r$). Second, we derive the regulator's mixed strategy that makes the firm indifferent between playing two of these strategies, and for which parameters the firm is better off not deviating from the resulting mix.

In any mixed-strategy equilibrium, each player randomizes over some actions to make the other player indifferent between their selected strategies. We first find that the probabilities p_d and p_r that make the regulator indifferent between a_d and a_r. These probabilities are:

$$\begin{aligned} p_r &\in \left[0, \frac{\pi_r}{\pi_d + \pi_r}\right]; \\ p_d &= \frac{\pi_d - (1 - p_r)\,\pi_r}{\pi_d}; \\ 1 - p_d - p_r &= (1 - p_r)\,\frac{\pi_r}{\pi_d} - p_r. \end{aligned}$$

Note that p_d is always positive so that the firm satisfies the disclosure requirement with strictly positive probability. With the complementary probability, the firm plays either $(0, 0)$ (satisfying the readability requirement with zero probability), $(0, r)$ (always satisfying one or the other requirement), or both if she is indifferent between $(0, 0)$ and $(0, r)$.

The regulator has only two non-dominated strategies, a_d and a_r. Because the firm always plays $(d, 0)$ with positive probability, we consider next the mixed strategy the regulator adopts to render the firm indifferent between $(d, 0)$ and either $(0, r)$ or $(0, 0)$. We use $p^r_{a_d}$ and $p^0_{a_d}$ to denote the probabilities of playing (a_d) that make the firm indifferent

between $(d,0)$ and $(0,r)$ or $(0,0)$, respectively:

$$\begin{aligned} p^r_{a_d} &= \frac{\pi_r}{\pi_d + \pi_r}; \\ p^0_{a_d} &= \frac{(1-\pi_r)\,k}{\pi_d v - \pi_r k}. \end{aligned}$$

With these probabilities, we characterize all possible equilibria. First, given the regulator's strategy $p^0_{a_d}$ or $p^r_{a_d}$, we find the expected payoffs of the firm playing $(d,0)$ and $(0,r)$ or $(0,0)$, respectively. Then, it suffices to identify the parametric values such that the other strategies are dominated given the regulator's strategy.

Proposition 2 (Constrained Regulator)**.** *Suppose the regulator is constrained and can inspect a policy for only one requirement. Let*

$$\underline{k} := \frac{\pi_r \pi_d}{\pi_r + \pi_d + \pi_r \pi_d} v < \frac{\pi_r \pi_d}{\pi_r + \pi_d - \pi_r \pi_d} v =: \overline{k}.$$

Then the following equilibria exist:

1. *if $\frac{1}{2} < \pi_r < \pi_d < 1$ and $0 < k < \underline{k}$, there is a continuum of payoff-equivalent equilibria in which the regulator mixes between a_d and a_r and the firm complies with both requirements with probability one;*

2. *if $\frac{1}{2} < \pi_r < \min\left(\pi_d, \frac{\pi_d}{3\pi_d - 1}\right) \leq \frac{2}{3}$ and $\underline{k} < k < \overline{k}$, in the unique equilibrium the regulator mixes following $p^r_{a_d}$, and the firm complies with either the content or the readability requirement (but not both) according to $p_d = \frac{\pi_d}{\pi_d + \pi_r}$;*

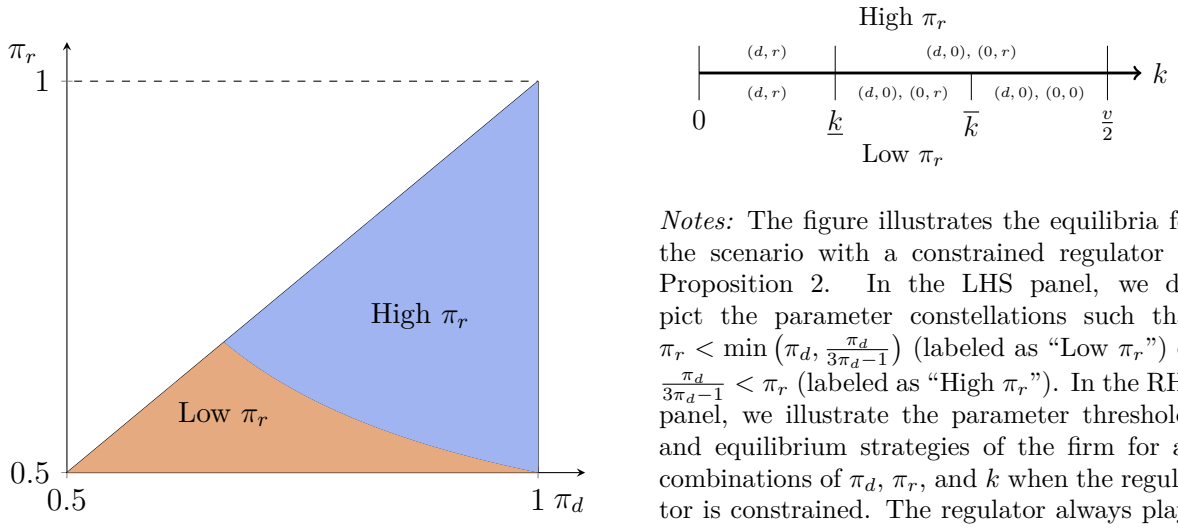
3. *if $\frac{1}{2} < \pi_r < \min\left(\pi_d, \frac{\pi_d}{3\pi_d - 1}\right) \leq \frac{2}{3}$ and $\overline{k} < k < \frac{v}{2}$, in the unique equilibrium the regulator mixes following $p^0_{a_d}$, and the firm either complies with the content requirement or does not comply with either of the requirements according to $p_d = 1 - \frac{\pi_d}{\pi_r}$;*

4. *if $\frac{\pi_d}{3\pi_d - 1} < \pi_r < \pi_d < 1$ and $\underline{k} < k < \frac{v}{2}$, in the unique equilibrium the regulator mixes following $p^r_{a_d}$, and the firm complies with either the content or the readability requirement according to $p_d = \frac{\pi_d}{\pi_d + \pi_r}$.*

Proof. The proof is relegated to Appendix A. □

Figure 1 illustrates the parameter space and the respective equilibria. Proposition 2 states that depending on k, the game either has a unique equilibrium for $k \geq \underline{k}$ or has a continuum of payoff-equivalent equilibria for $k < \underline{k}$. This multiplicity arises because the cheaper it is to comply with the requirements, the easier it is for the regulator to induce the firm to play (d,r).

Figure 1: Equilibria for Constrained Regulator



Notes: The figure illustrates the equilibria for the scenario with a constrained regulator in Proposition 2. In the LHS panel, we depict the parameter constellations such that $\pi_r < \min\left(\pi_d, \frac{\pi_d}{3\pi_d - 1}\right)$ (labeled as "Low π_r") or $\frac{\pi_d}{3\pi_d - 1} < \pi_r$ (labeled as "High π_r"). In the RHS panel, we illustrate the parameter thresholds and equilibrium strategies of the firm for all combinations of π_d, π_r, and k when the regulator is constrained. The regulator always plays a mixed strategy.

The latter collection of equilibria leads to the same outcome: The regulator inspects either the disclosure or the readability requirement with positive probability in a way that makes the firm strictly better off complying with both than deviating to any other strategy. The easier it is to comply (lower k), the wider the range of mixed strategies that lead to this outcome. When $k = \underline{k}$, strategy $p^r_{a,d} = \frac{\pi_r}{\pi_d + \pi_r}$ makes the firm play (d, r) with probability one and generates the only mixed-strategy equilibrium. In contrast, as k approaches 0, any $p^r_{a,d} \in (0, 1)$ leads to the same outcome.[18]

4.3 Varying the Regulator's Resources

Propositions 1 and 2 reflect the ability of the regulator to induce compliance through inspection of the disclosure and readability requirements. Recall that k^u denotes the threshold below which an unconstrained regulator can induce full compliance from the firm. This value satisfies:

$$0 < \underline{k} < k^u < \overline{k} < \frac{v}{2}. \tag{1}$$

The above ordering reveals that an unconstrained regulator can induce full compliance for a more extensive set of values k. This is because if $\underline{k} < k < k^u$, a constrained

[18]Proposition 2 accounts for all equilibria except for knife edge scenarios in which the agents are indifferent between two of the above equilibria. It is not difficult to show that no other mixed-strategy equilibria exist except those characterized in Proposition 2. First, suppose a mixed-strategy equilibrium exists that involves the firm playing (d, r) with some positive probability different from one. Then, the regulator will optimally reply playing the best response to the other action with probability one, to which the firm's best response is something other than (d, r) as illustrated above. Further, there is no fully mixed-strategy equilibrium in which the firm plays only $(0, 0)$ and $(0, r)$, as there is no mixed strategy employed by the firm that makes the regulator indifferent between a_d and a_r where $p_d = 0$.

regulator is unable to induce full compliance no matter the relative value of π_r and π_d (see Proposition 2). Similarly, equilibrium compliance is lower for very high levels of compliance costs.

Proposition 3. *Suppose a tightening of the regulator's budget that renders a once unconstrained regulator constrained. With a now constrained regulator, compliance with the disclosure and readability requirements is weakly lower in equilibrium and strictly lower if $\underline{k} < k < k^u$ or $\overline{k} < k < \frac{v}{2}$. If $k > k^u$, the firm complies with at most one requirement: only d if the regulator is unconstrained, either d or r (or neither) if the regulator is constrained.*

Proof. The proof is relegated to Appendix A. □

In equilibrium, the firm never focuses on readability more than on disclosure. If the regulator can inspect both requirements, the firm either complies fully or ignores readability depending on how costly compliance is. If the regulator is constrained, instead, the firm fully complies only if it is very cheap to do so (low k). Otherwise, she invests in disclosure or readability, and focuses relatively more on disclosure, if costs of compliance are intermediate and readability is not easily enforced (that is, if π_r is relatively low). If costs are very high and readability is not easily enforced, the firm either focuses on disclosure or chooses not to comply with either requirement. If readability is easily enforced (that is, if π_r is relatively high), the firm once again invests in disclosure or readability, and relatively more on disclosure.

4.4 Discussion of Results

Our model yields several empirical predictions. First, more stringent regulation (i.e., regulation of the disclosure and readability requirements) weakly encourages compliance. To see why, suppose that the regulator in our framework cannot audit any requirements.[19] The firm will only ever choose not to comply with requirements that cannot be audited. In this sense, more stringent regulation being introduced will have a positive effect on compliance with both disclosure and readability if regulators act optimally in their role of auditors, and a positive effect on disclosure alone if they do not.

Prediction 1. *Privacy policies become longer and, to a lesser extent, more readable after the introduction of more stringent regulation on disclosure and readability requirements for privacy policies.*

The model highlights the three factors that govern which of the arising equilibria we ought to expect: cost of compliance k, enforceability π_j, and the regulator's budget constraint. While firms generally tend to comply with disclosure in equilibrium, these

[19] In terms of the model, this regulator could only play action a_0.

three factors determine the environments in which we will see relatively more compliance with the readability requirement.

The cost of compliance and enforceability are closely related. On one hand, for low costs of compliance, the predicted level of compliance for all π_j and levels of constraint of the regulator is higher. On the other hand, the higher the perceived risk of scrutiny by regulators, the more we expect compliance to arise. We expect firms that anticipate more thorough regulatory scrutiny or for whom drafting legal documents is relatively cheap to comply more with both requirements.

Prediction 2. *With more stringent regulation, larger firms draft more readable privacy policies relative to smaller ones; similarly, firms operating in markets subject to stricter scrutiny draft more readable privacy policies compared to firms that do not.*

We expect firm size to be a good predictor of compliance: larger firms tend to have lower effective drafting costs compared to a smaller firm[20], and to be more likely subject of auditing given their prominence.

On the other hand, some markets are likely to attract more attraction than others when it comes to regulatory audits. Firms active in markets in which more data tends to be collected and processed, for example, would be more likely to be the object of scrutiny. A more concentrated market, with fewer active firms, should also attract more regulatory attention than a dispersed market with many smaller actors. The interaction between a firm's size, scope, and competitive environment is not obvious. We bring the question of the role of these different dimensions to the data and provide a detailed discussion in Section 8.

The model shows that the budget constraint of regulators plays an important role in their ability to incentivize compliance. Intuitively, the more resources are available to a regulator, the more thorough he can be in his audits. This thoroughness translates to different effects for the two requirements when interacting with the cost of compliance and perceived level of scrutiny. In general, however, a firm facing scrutiny by an unconstrained regulator always complies more with both requirements than a firm facing a constrained regulator:

Prediction 3. *With more stringent regulation, firms operating in jurisdictions with less budget-constrained regulators draft more readable privacy policies than firms operating in the jurisdiction of more constrained regulators. The same firms also draft longer policies, but the effect on disclosure is smaller compared to that on readability.*

Understanding the role of regulators' constraints on compliance is important beyond the specific predictions of this model. It is crucial to understand under which conditions

[20]Larger firms generally have access to sizable legal teams competent in the fields relevant to the firms' operations. Therefore, a larger firm would be able to draft a complying privacy policy more easily because of their in-house expertise.

enforcement of legal requirements can be carried out effectively and efficiently. The limitations that come with reductions in the budget of regulatory agencies affect enforceability, and therefore compliance, more generally and beyond the context of the GDPR.

Finally, our equilibrium analysis reveals that a regulator facing budget constraints should focus more on readability than on disclosure in its audits. Suppose, however, that a constrained regulator were not to inspect readability at all, perhaps because of how difficult it is to properly evaluate it. Anticipating this, firms would optimally disregard the readability requirement and comply only with disclosure. Because an unconstrained regulator would always inspect both requirements, increasing the budget of the regulatory agency would lead to a higher level of compliance with the readability requirement. However, since firms with relatively high costs of compliance never comply with both requirements when facing an unconstrained regulator, the level of compliance with disclosure might actually decrease when the regulator has a larger budget at his disposal.

5 Data

For our empirical analysis, we construct an unbalanced quarterly panel with the texts of some 580,000 privacy policies posted by some 75,000 firms between 2014 and 2021. We complement this information with firm, industry, and state-level characteristics, that is, budget information for state data protection authorities (DPA). In this section, we describe our data sources and the construction of the privacy policy panel. We discuss our measures for disclosure and readability in the next section.

5.1 Privacy Policy Panel

We use an unbalanced quarterly panel of the texts of privacy policies of German firms posted between 2014 and 2021. We constructed the panel by web-scraping the Internet Archive (via the Wayback Machine) to obtain the historic versions of the policies using the following multi-step approach.[21]

1. We obtain approximately 570,000 URLs (uniform resource locators) of webpages from the 2019 wave of the Mannheim Web Panel (Kinne and Axenbeck, 2019) with URLs containing the term "datenschutz" (the German word for data protection) or "privacy". For each URL, we obtain a unique firm identifier for the Mannheim Enterprise Panel. We also extract from each URL the domain and top-level domain. From the complete set of URLs, we obtain the eight most common URL patterns as combinations of a given firm's domain, top-level domain, second-level domain, and subdirectory.

[21] We provide a detailed description of each of these steps in Appendix D.

2. We search the Internet Archive for each constructed URL, beginning with the URL from the 2019 wave. For all available pages of a given URL, we download and store the full HTML page and record the respective date. If, in a given quarter, a page is not found for that specific URL, the scraper circles through the list of common-pattern URLs. If this second step does not recover a page, we set the observation to missing. If, in a given quarter, we find multiple pages in the Internet Archive, we store the first page recorded (by date) in that quarter.

3. From each downloaded page (in HTML format), we extract the text of the respective privacy policy. We use a simple parser (that we manually calibrated and optimized using a viewer app) to capture the relevant text portions while ignoring other portions of the HTML pages (such as headers, pictures, or external links). We also delete empty pages or error pages from our sample.

4. For the construction of our final estimation sample, we impose a number of additional restrictions. First, we consider only German-language policies posted between Q1 2014 and Q2 2021. To further eliminate pages that are likely too short to contain privacy policies or too long to contain the privacy policies but nothing else, we drop observations that are shorter than the 2nd percentile and longer than the 98th percentile (measured in simple word tokens). Moreover, to ensure observations over the entire sample period (and to partially balance our panel), we restrict our sample to policies by firms for which we observe at least one observation (1) prior to the enforcement of the GDPR (May 25, 2018), and (2) after the enforcement of the GDPR.[22]

Our final sample comprises 585,329 privacy policies by 75,683 firms from Q1 2014 to Q2 2021. In Figure 2, we show the number of observations per quarter (i.e., the number of privacy policies by as many firms). For Q2 2016 and Q2 2018, we observe policies in two separate phases (indicated by the color shading). The average number of observations per firm is 4.4 pre-GDPR enforcement and 3.3 post-GDPR enforcement.

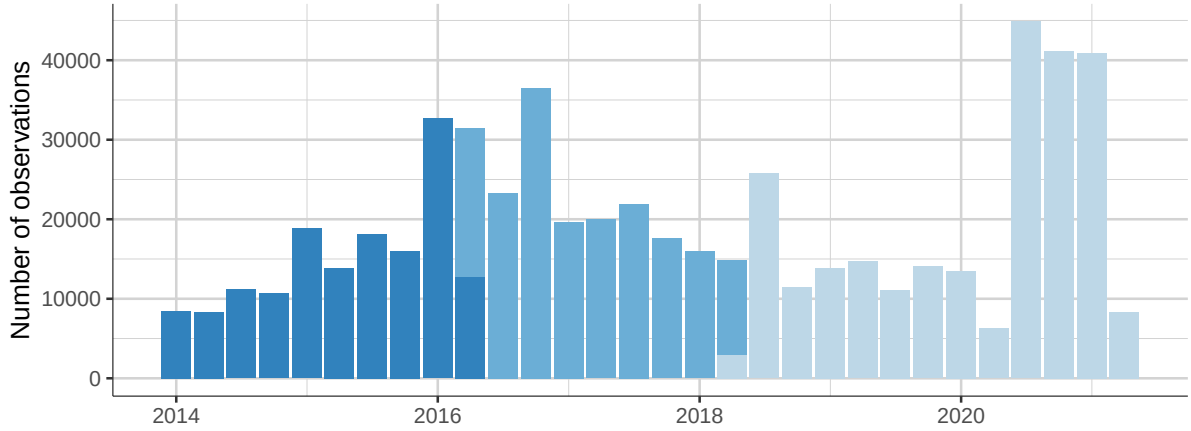
5.2 Firm and Industry-Level Characteristics

We complement the privacy policy panel with information on firms' employees and sales from the Mannheim Enterprise Panel (*Mannheimer Unternehmenspanel*, MUP). The Panel is the most comprehensive micro-data base of companies in Germany beside official administrative data (Bersch et al., 2014).[23] We report sample characteristics in Table 2.

[22] Johnson (forthcoming) argues that, in addition to the 2018 enforcement, the passage of the GDPR in May 2016 also needs to be considered when examining GDPR effects. As we will show later, we do not observe any effects around the 2016 passage or the GDPR.

[23] The MUP contains a large share of all companies in Germany, including micro-enterprises and self-employed freelancers (albeit underrepresented). It contains all companies in Germany that are

Figure 2: Observations by Quarter



Notes: The figure presents the number of observations (one policy per firm) per quarter for our estimation sample (Q1 2014 to Q2 2021). The different shades of blue indicate three time phases: pre-GDPR passage (May 4, 2016; Q2 2016), pre-GDPR enforcement (May 25, 2018, Q2 2018), and post-GDPR enforcement. In Q2 2016, and Q2 2018, we have observations in two phases (before and after the respective cut-off dates).

The average firm in our sample has 36 employees and sales of 15 million Euros. Following official reporting standards, we define micro firms as those with less than 10 employees, small and medium-sized firms (SMEs) with between 10 and 250 employees, and large firms with more than 250 employees. Out of these sub-samples of firms, for which we have employment numbers for at least one observation (65,863 firms), 61.6% are micro firms (80% in 2017 MUP), 36.3% are small and medium-sized enterprises (19.1% in 2017 MUP), and 2% are large firms (0.9% in 2017 MUP). In our sample, micro and large firms are underrepresented and SMEs are overrepresented, relative to the Mannheim Enterprise Panel in 2017.

We further obtain from the MUP the four-digit NACE Rev. 2 codes of the industry of firms' primary business activities. Our largest sector is services with 58.6% of all firms in 2017 (55.8% in the MUP), followed by trade with 22.3% (23.9% MUP), manufacturing with 9.6% (6.7% in MUP), construction with 7.0% (10.7% MUP), utilities with 1.5% (0.9% MUP), and agriculture/mining with 1% (2% in MUP). In our estimation sample, services, manufacturing, and utilities are over-represented whereas trade, construction, and agriculture/mining are underrepresented.

Last, we calculate annual numbers for the Herfindahl-Hirschman Index (HHI) for all 4-digit NACE industries using firm-level sales information from the Mannheim Enterprise

economically active to a sufficient degree. However, detailed information is not available for all firms; small firms are generally underrepresented. The reason for this bias lies in reporting rules in Germany. Medium sized and large firms are obligated to report balance sheet information, a statement of income, and notes on the accounts. Small firms are exempt from the reporting requirements of a statement of income.

Table 2: Sample Characteristics

	Obs.	Mean	Std.	Min	Max
Number of observations per firm	75683	7.734	4.67	2	30
… in pre-GDPR enforcement phase	75683	4.446	3.69	1	18
… in post-GDPR enforcement phase	75683	3.288	2.17	1	13
Employees (firm-level means)	65863	36.446	408.48	1	48300
… Micro	40578	3.72	2.54	1	10
… Small and medium-sized (SME)	23920	39.222	42.13	10	249.6
… Large	1365	960.678	2671.81	250	48300
Sales (in million; firm-level means)	55656	14.942	351.78	0	62379.6
Herfindahl-Hirschman Index (HHI; in 2017)	44883	551.131	1178.23	1.5	10000
Economic Sector (2017)	**Estimation sample**		**MUP**		
Agriculture/Mining	688	1.03%	1.96%		
Manufacturing	6387	9.56%	6.72%		
Utilities	1028	1.54%	0.92%		
Construction	4679	7.01%	10.69%		
Trade	14907	22.32%	23.89%		
Services	39105	58.55%	55.82%		
	66794				

Notes: We report sample size and firm-level characteristics for the estimation sample. The number of employees and sales figures (firm-level means) are from the Mannheim Enterprise Panel (MUP), waves 47 to 61. Small firms have less than 10 employees; small and medium-sized enterprises (SME) have between 10 and 250 employees; large firms have 250 employees or more. The reported numbers are the averages of all of a given firm's observations. Market concentration information (Herfindahl-Hirschman Index) is calculated from the Mannheim Enterprise Panel for 2017 (using the four-digit NACE industry classification). Economic sectors are based on a firm's primary NACE Rev. 2 code (as reported in 2017): Agriculture are sections A and B; manufacturing is section C; utilities is sections D and E; construction is section F; trade is section G; and services is sections H, I, J, K, L, M, N, P, Q, R, and S.

Panel.

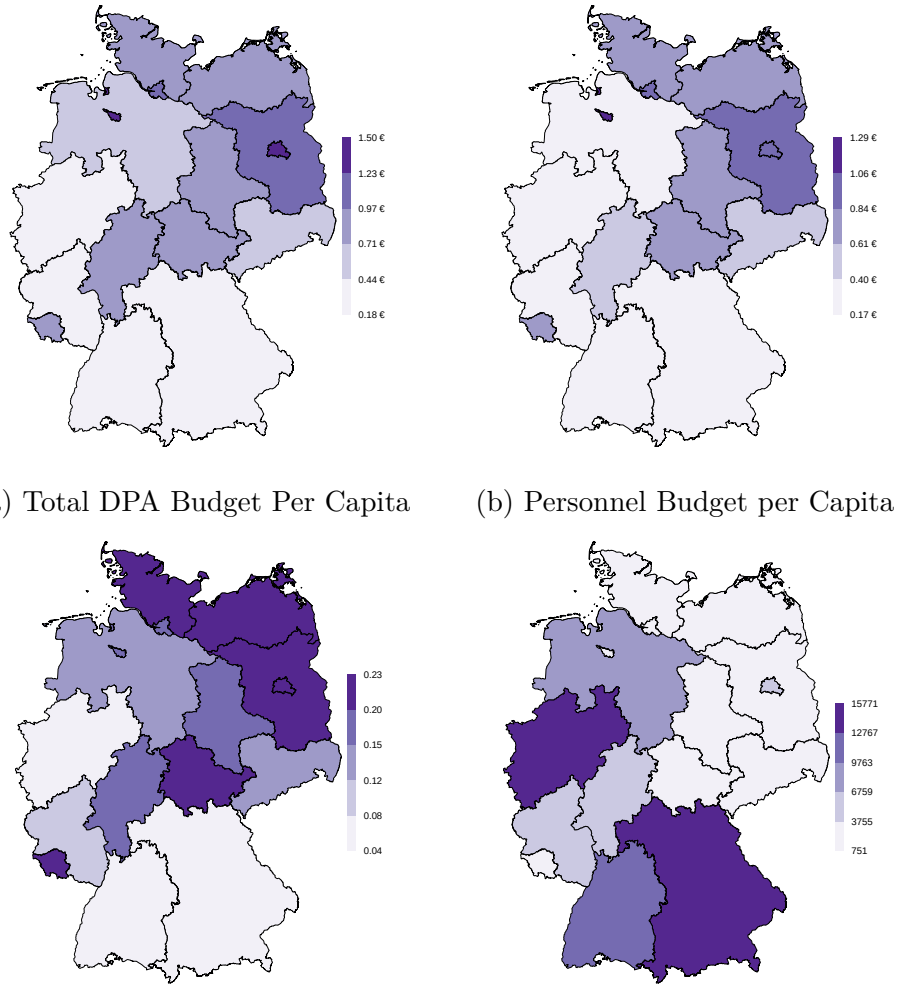
5.3 Budgets of German State Data Protection Authorities

Germany uses a federal system for data protection regulation. Each state has its own data protection authority (DPA) that regulates the compliance of firms located in the respective state. We use the official budget of the DPA in the firm's home state to proxy regulatory capacity: an authority with more resources is less likely to be a constrained regulator in the sense of our theoretical model.

The MUP provides firm-level address information (firm's headquarters) for the years 2013 through 2019. For observations in 2020 and 2021, we use the address information available for 2019. We further extrapolate forward and backward, using a firm's first address information for all prior observations and a firm's last address information for all following observations. We also interpolate missing observations between two observations for which the firm's state has not changed. If, between two observations, the firm has moved to another state, we do not fill the in-between observation (missing values).

We obtain state-level budget information from state governments' website. We collect information on a DPA's overall budget (i.e., budgeted expenditures) and (budgeted)

Figure 3: Budgets of German State Data Protection Authorities



(a) Total DPA Budget Per Capita

(b) Personnel Budget per Capita

(c) Share of Overall State Budget

(d) Number of Firms per State

Notes: This figure presents (a) the budgeted total expenditure per capita, (b) the budgeted personnel/labor expenditure per capita, (c) the total DPA budget over a state's total expenditure per capita (reported in '000), and (d) the total number of firms in our sample (by state). All values are for the year 2018. See Figures A.1, A.2, and A.3 for state-level time series for the budget-related numbers.

expenditures for labor/personnel.[24] We further collect state-level population information from the Federal Statistical Office of Germany to construct total and labor budget per capita.[25] Schütz (2018) argues that budget (or staff) by inhabitant may be misleading because of the fixed costs incurred by DPAs. The result is upward bias in particular for smaller states. He then concludes that a DPA's budget over the respective state's total budget provides a different perspective of the relative importance of the data protection agency within the state. To this end, we collect states' annual expenditures to construct a state's data protection budget as share of the state's overall budget.

[24]We consider only the original budget numbers and disregard any revised budgets.

[25]Population by state is in Data Table 12411-0010, available at https://www-genesis.destatis.de/genesis//online?operation=table&code=12411-0010.

Figure 3 summarizes the budget situation of German DPAs for 2018. We plot the (total and labor) budget per capita (in Euros) and the relative DPA budget (over state's total budget) for all 16 states. As observed by Schütz (2018), we see significantly higher per capita figures (in panels (a) and (b)) in small states (Berlin, Bremen, and Hamburg), but also in states in the northeast of the country. When considering the relative DPA budget (in panel (c)), small states no longer have a budgetary advantage. However, states in the northeast and the Saarland (in the west-southwest) have better equipped DPAs. In the Appendix, we plot the time series for each of the three measures (Figures A.1, A.2, and A.3).

6 Measuring Disclosure and Readability

We describe the construction of our proxies for disclosure (Art. 13–14 GDPR) and readability (Art. 12 GDPR).

6.1 Simple Measures of Informational Volume

We begin our section on data construction with a few simple measures of informational volume. We can generally expect more information conveyed in privacy policies to translate into longer texts with a larger number of distinct terms. To capture longer texts, we construct total word counts, counts of unique words,[26] and sentence counts. In Panel (a) in Table 3, we report the sample statistics of these simple measures. The average policy has a total count of 917 words with 420 distinct (single-word) terms in 84 sentences. We see a considerable amount of variation over our sample period, where the shortest policy uses 22 unique words (61 total words in 2 sentences) and the longest policy comprises an extensive lexicon with more than 2000 unique words (more ten 5600 total words in 569 sentences).

We further construct a measure of the number of distinct topics (or themes) covered in a privacy policy. To that end, we determine the "main topic" of a given paragraph and then tally the number of distinct main topics for each policy. We use probabilistic topic models to help us discover the topics in our policy corpus. More specifically, we apply the *Latent Dirichlet Allocation* (LDA) model (Blei et al., 2003), which Blei and Lafferty (2009) describe as the "simplest topic model" and that "has proven hugely popular" (Taddy, 2012).[27] Probabilistic topic models uncover the latent topical structure of a document by analyzing the co-occurrence of tokens (i.e., words, terms, or phrases) used

[26]For counts of total and unique words we first pre-process our text corpus (convert to lower case; remove punctuation, numbers, common words ("stopwords"), single letters, and roman numerals) and then stem all words.

[27]Blei and Lafferty (2009), Blei (2012), or Steyvers and Griffiths (2011) provide an introduction to probabilistic topic models. We use a standard implementation of LDA (Grün and Hornik, 2011).

Table 3: Text Volume and Disclosure

	Mean	Std.	Min	Max
Panel (a): Informational volume				
Total word count	917.21	839.2	61	5614
Unique words	419.55	270.93	22	2039
Number of sentences	84.24	75.18	2	569
Distinct topics	12.36	9.66	1	47.1
Panel (b): Disclosure				
Disclosed words	660.23	678.85	0	5614
Topic-weighted words	1006.42	1078.84	0	10998.2

Notes: This table reports basic text characteristics capturing both informational volume and disclosure in the 585,329 privacy policies in our estimation sample. We use the `tm` package for the construction of the total word count and the number of unique words (after basic text pre-processing steps); we use the `topicmodels` package to calculate topic models; we use the `quanteda.textstats` package (Benoit et al., 2018) to count the number of sentences. The number for topic models are based on 10 iterations of random draw of 5,000 policies.

in the document. The underlying idea is that authors first decide which topics to cover before drafting the document. A document thus becomes a collection of multiple topics.[28]

We follow a two-step approach. First, to obtain the main topic for a given paragraph, we estimate the topic model with $K = 50$ topics on the corpus of paragraphs.[29] For each paragraph c in a policy d, we obtain the per-document topic distribution $\vec{\theta}_{cd} = \left(\theta_{1|cd}, \ldots, \theta_{K|cd}\right)$ over K topics, with $\sum_{k=1}^{K} \theta_{k|cd} = 1$. Each $\theta_{k|cd}$ represents the weight with which a topic k is covered in a paragraph c. Second, we assume that, in practice, each paragraph was written to cover a single topic. We call this topic the *main topic* k_{cd} of a paragraph and define it as the topic with the highest topic density $\theta_{k|cd}$, so that $k_{cd} = \arg\max_{k=1,\ldots,K} \theta_{k|cd}$. For a count of main topics for each policy, the union of main topics gives us the set of distinct topics that are a main topic for at least one paragraph.[30]

Table 3 reports the results from this approach. The average policies covers around 12 distinct topics. The shortest policy is quite minimalist (one distinct topic) whereas the

[28]The LDA topic model describes such a topic k as a *per-topic word distribution* $\vec{\beta}_k$ over the vocabulary of N tokens (i.e., a $(1 \times N)$ vector for each topic k). Moreover, for our corpus of privacy policies, say $\mathcal{D}$, holding documents that cover K topics, each document $d \in \mathcal{D}$ will exhibit these K topics with different proportions according to a *per-document topic distribution* $\vec{\theta}_d$. The data we observe are the documents in a collection $\mathcal{D}$ and the tokens $\vec{w}_d$ used in each document. The topics, however, are not observed. We apply LDA to reverse this process of topic generation and automatically discover the latent topical structure. This means that we obtain estimates for $\vec{\beta}_k$ (for $k = 1, \ldots, K$) and $\vec{\theta}_d$ (for $d \in \mathcal{D}$). In Figure A.4 (taken from Ganglmair and Wardlaw (2017)), we provide a stylized depiction of this process for two documents ($d = 1, 2$) and three topics ($k = 1, 2, 3$).

[29]Estimating the topic models on a corpus of shorter documents (i.e., paragraphs) follows Brody and Elhadad (2010). We first perform standard pre-processing steps and define tokens as unigrams. Estimating topic models is computationally intensive. We limit the number of tokens to 3,000 (by frequency) and estimate our topic model on paragraphs of 5,000 randomly drawn policies, predicting the topic assignments for *all* paragraphs. The number of topics K is chosen to balance the additional granularity of higher K with the computational burden.

[30]We obtain a $(1 \times K)$ vector $\vec{k}_d$ with $K = 50$ elements, each being equal to 1 if topic k is a main topic at least once, and zero otherwise. The number of main topics is then $\sum_{k=1}^{K} \vec{k}_d$.

paragraphs in the longest policy contain 47 distinct topics. We use the these topic model results for the construction of our disclosure proxy in the next section.

6.2 Disclosure

To evaluate firms' compliance with the disclosure requirements in the GDPR, we want to identify those parts of a privacy policy that contain the required information (e.g., by Art. 13 and 14). We use the results from the LDA topic models to identify paragraphs that are more or less likely to contain information related to a firm's disclosure. Using higher weights for the word counts of more relevant paragraphs (those more likely disclosing relevant information) and lower weights for those of less relevant paragraphs, we thus construct a measure of the *topic-weighted informational volume* as our proxy for a firm's disclosure. We take a multi-step approach:

1. From the per-paragraph assignments of main topics, we calculate a topic distribution where Θ_k represents the fraction of paragraphs with topic k as their main topic.

2. We identify all paragraphs (both pre-GDPR and post-GDPR) that contain information related to disclosures per Art. 13 and 14, using simple text parsing techniques.[31] The total word counts of disclosing paragraphs is the total number of *disclosed words*. For the subset of disclosing paragraphs taken from post-GDPR policies, we calculate the topic distribution with respective densities $\tilde{\Theta}_k$. Using the main-topic distributions for all paragraphs (Step 1) and for disclosing paragraphs, we calculate a topic weight factor $\phi_k = \frac{\tilde{\Theta}_k}{\Theta_k}$ for each k. We interpret a topic k with $\phi_k > 1$) (or $\tilde{\Theta}_k > \Theta_k$) as one that is more likely capturing information required by Art. 13 and 14 than an alternative topic k' with $\phi_{k'} < 1$.

3. We obtain the word count $w_{c|k}$ for each paragraph c of a given main topic k.[32] We multiply the paragraph word counts by the paragraph's respective topic weight factor to obtain the number of *topic-weighted words* ($\sum_c \phi_k w_{c|k}$) as our measure of disclosure. Table A.2 illustrates this construction of topic-weighted information volume using two simple examples.

We provide descriptive statistics for our disclosure proxies in Panel (b) in Table 3. For the average policy, a bit more than two out of three words are in disclosed words (Step 2) in disclosing paragraphs (with the longest policy holding all of its words in disclosing paragraphs). Because the list of terms and phrases we use to identify disclosing

[31] Table A.1 provides the regex expressions we use.

[32] The total word count of the policy, as reported in Table 3, is only an approximation of the sum of paragraph word counts, because we drop paragraphs that are very short and therefore likely text stubs (such as titles).

Table 4: Readability Factors and Scores

	Mean	Std.	Min	Max
Panel (a): Readability factors				
Word length (in syllables)	2.16	0.07	1.4	3.4
Sentence length (in words)	17.84	3.26	4.2	222
Share of big words (5+ syllables)	0.21	0.04	0	0.5
Panel (b): Readability scores				
German Flesch Reading-Ease score	35.98	5.64	-185.8	86.7
LIW	56.13	3.94	22	260.3

Notes: This table reports both components used to calculate readability scores and descriptive statistics for the German FRE and LIW score for all 585,329 privacy policies in our estimation sample. We construct the word length, sentence length, German FRE, the LIW using the `quanteda.textstats` package (Benoit et al., 2018). For the share of big words, we apply the syllable counter in the `sylly` package (with German language support in `sylly.de`) to the term lists (from `tm`), count the number of words with five or more syllables, and divide by the total word count.

paragraphs is biased toward post-GDPR language, we do find policies with no disclosed words. As for all other measures of text volume and disclosure, we observe significant heterogeneity of topic-weighted words (Step 3) across policies.

6.3 Readability

6.3.1 Text Structure

Many factors determine how readers comprehend written texts. For example, the use of common words will make texts more accessible to a wider audience, whereas the use of specialized terms or jargon will render texts more difficult to understand. Similarly, shorter sentences or simpler and shorter words will increase the readability of a text and transparency of its content.[33] In Table 4, we report three factors from our sample of privacy policies: average *word length* (in syllables), average *sentence length* (in words), and the share of *big words*, defined as words with at least five syllables. We interpret texts with longer words, longer sentences, and with of more big words as less readable or accessible.

To allow for better comparison, we report in Table A.3 readability factors for a number of different text corpora. Texts that are expected to be simpler also exhibit simpler factors. For instance, for a sample of plain-language (or simple-language) news articles published weekly by *Deutschlandfunk* (a public-broadcasting radio station in Germany), we find readability factors that are significantly lower than those for the sample of privacy

[33]The plain language movement, built on the work by Mellinkoff (1963), advocates the use of *plain language* or *simple language*, particularly by governmental bodies and public agencies, to provide for more inclusive communication and information. In Germany, for instance, this culminated in the inclusion of the movement's goals in the Equality for Persons with Disability Act (BGG) of 2002 that stipulates barrier free access to communication and information, such access being ensured by the use of plain language ((Kellermann, 2014; Stefanowitsch, 2014).

policies. The official text of the German version of the GDPR, on the other hand, uses longer words and longer sentences.[34]

6.3.2 Readability Testing

Readability scores and indices are developed to assess the reading ease (or difficulty) of texts. Such scores have been used (e.g., in the United States) in regulatory contexts.[35] They are typically constructed as weighted averages of a set of different readability factors (including the three factors summarized above), and make it easier to interpret the output. The literature, however, knows a large number of scores and indices developed for different languages and purposes that also vary in their popularity and use.[36]

Legal writings (such as privacy policies), however, comprise a special text category,[37] and it is far from obvious which index is the most suitable for the purpose of analyzing the readability of privacy policies. For our main analyses, we use two scores: First, we use the German version of the Flesch Reading Ease Score (Flesch, 1948; Amstad, 1978) because of its established use in regulatory context.[38] It is defined as

$$\text{German FRE} = 180 - ASL - (58.5 \times AWL) \tag{2}$$

where ASL and AWL denote the average sentence and average word length (in syllables), respectively. Higher values indicate better readability.

For our second readability measure, we take a data-driven approach. We follow Benoit et al. (2019) who evaluate the textual complexity in texts (i.e., political communication) by fitting a domain-specific measure of textual sophistication. The authors use the Bradley-Terry model for pair-wise comparisons (Bradley and Terry, 1952). To implement this approach, we first hand-collect about 4,000 pair-wise comparisons of portions of privacy policies (taken from our sample), asking subjects to rank the two text snippets in a given pair by their readability.[39] These pair-wise comparisons align well with the ranking

[34] We also compare privacy policies with speeches and statements by former German Chancellor Angela Merkel and decisions issued by the German Constitutional Court. The picture is mixed. For instance, speeches use on average shorter words but also more big words. Wikipedia pages, on the other hand, exhibit longer sentences than privacy policies, but use fewer big words.

[35] In Michigan and Massachusetts, an insurance contract must have a Flesch Reading Ease (FRE) Score of at least 50 (Michigan Compiled Laws, Section 500.2236 (2020); General Laws of Massachusetts, Title XXII, Chapter 175 Section 2B. (2014)); in Texas, the minimum score of the FRE is 40 (Texas Insurance Code, Section 2301.053 (2019)) (Wagner, 2023). Similar guidelines (with a minimum score of 45) exist in Florida (Florida Statute §627.4145, Readable language in insurance policies; available at https://flsenate.gov/Laws/Statutes/2021/0627.4145).

[36] See Table A.4 for a list of readability scores and their use in the literature.

[37] They have unusually long sentences with an average sentence containing twice as many words as in other categories of texts (Gustafsson, 1984).

[38] For academic work, see Lin and Osnabrügge (2018) or Wojahn et al. (2015).

[39] We used the results from our LDA model to identify paragraphs that are central to understanding the processing of personal data. We then selected a random sample of paragraphs each 60–80 words long (ruling out multiple paragraphs from the same firm) and constructed 700 text pairs. In batches of

of text snippets using, for instance, average word length or sentence length.[40]

The so-collected sample serves as our "gold standard" for determining text readability in the next step in which we apply an unstructured Bradley-Terry model (Bradley and Terry, 1952). This model was originally developed for sports competition to rank the best contestant. It estimates the odds that a contestant will outperform another in a competition, thus ranking comparisons according to their "ability." The model's output are the relative abilities resulting from the pairwise comparisons, with higher values indicating higher abilities. Applying this to the textual sophistication of texts, excerpts of privacy policies can also be ranked according to their "ability," in which texts compete for readability. By applying the Bradley-Terry model, each text snippet is ranked according to its relative readability. Furthermore, for every text snippet we compute the scores of common readability indices in order to connect the Bradley-Terry outcome with different readability scores. Following Benoit et al. (2019), we then run a number of random-forest regression models to find the readability index that best explains the outcomes of our pair-wise comparisons, using the abilities of the Bradley-Terry model as input features.[41] The higher the increase in node purity, the better an index is at predicting pairwise comparisons. The best readability score (as best predictor of the data) is the *läsbarhetsindex* (LIW) (Björnson, 1968):

$$\text{LIW} = ASL + \frac{100 \times n_{wsy \geq 7}}{n_w} \tag{3}$$

with ASL the average sentence length (in words), $n_{wsy \geq 7}$ the number of words with at least seven syllables, and n_w the total number of words. Higher values indicate lower readability (unlike for the German FRE). Note that the patterns for both the German FRE and the LIW align well with our sample of pair-wise text comparisons.[42] While the LIW is not among the most popular readability scores in the literature (Courtis, 1995; Ezat, 2019), it outperforms all other scores typically used in research. It is of note that

100, we assigned the text pairs to 14 human subjects (recruited via Upwork). To evaluate each subject's performance, we added to each batch 10 pairs that contained two identical snippets and 10 pairs that matched another pair in that batch but with the text snippets in reverse order. We exclude the results from these additional 20 pairs from our final sample. Each batch was evaluated by at least six subjects.

[40] In Figure A.5, we show that as difference in average word length (or sentence length) of two snippets in a pair increases, the percentage of pairs for which the human assessment aligns with the score-based ranking increases.

[41] The random-forest model ranks each predictor, in this case represented by the readability score of a snippet, according to its *importance* and enables us to see how each index performs in predicting the outcome of our gold standard data. The importance of an index is determined by calculating the average increase in "node purity" when that variable is used to split the tree at a specific point, aiming to predict the outcome variable (the snippet's ability). Node purity is measured by the residual sum of squares, where a lower value indicates better purity. We plot the results (i.e., increase in node purity) for a list of 33 readability scores in Figure A.7.

[42] In Figure A.6, we show that as difference in the German FRE (RHS panel) and the LIW (LHS panel) increase, the percentage of pairs for which the human assessment aligns with the score-based ranking increases as well.

less popular scores tend to outperform the more popular ones.[43]

We provide summary statistics of the German FRE and LIW index for our estimation sample in panel (b) of Table 4. For easier interpretation, we also calculate the scores for other text corpora. We report the result in Table A.3. Simple-language news pages are easier to read than our privacy policies, and we find similar patterns for other text domains (political speeches, German constitutional court decisions, or Wikipedia pages). Ironically (yet in line with the literature on the European Commission's communication quality (Rauh, 2023)), the German text of the GDPR itself (the Datenschutzgrundverordnung/DS-GVO, a seven-syllable word) is highly unreadable, with a LIW score almost 50% higher than that for privacy policies.[44]

7 Results for Firms' GDPR Responses

In this section, we document GDPR-associated changes in the amount of disclosure in and the readability of privacy policies. Following our theoretical framework (Prediction 1) we conjecture that the increased stringency of the transparency requirement following the introduction of the GDPR in Q2 2018 leads to longer privacy policies that disclosure more information to users. We further conjecture that firms write better privacy policies that are easier to read for users. This latter effect, if it exists, is weaker than the effect on disclosure.

7.1 Disclosure Before and After the GDPR

In panels (a) and (b) of Figure 4, we plot the quarterly averages or our measures for informational volume (unique words, sentences, and topics) and disclosure (disclosed words and weighted words), relative to the respective values in Q1 2014 (provided in the titles of each figure).[45] An increase by one implies a doubling of the respective variable.

All graphs in the top two panels paint a similar picture: privacy policies have more than doubled in length, and disclosure has more than tripled, with the enforcement of the GDPR.[46] For quarter Q2 2018, we plot average values for before and after the enforcement of the GDPR; the documented gap is therefore within-quarter. All but the count of distinct topics continue to increase after Q2 2018. This suggests that the breadth of policies remains relatively constant, whereas the details of the documents

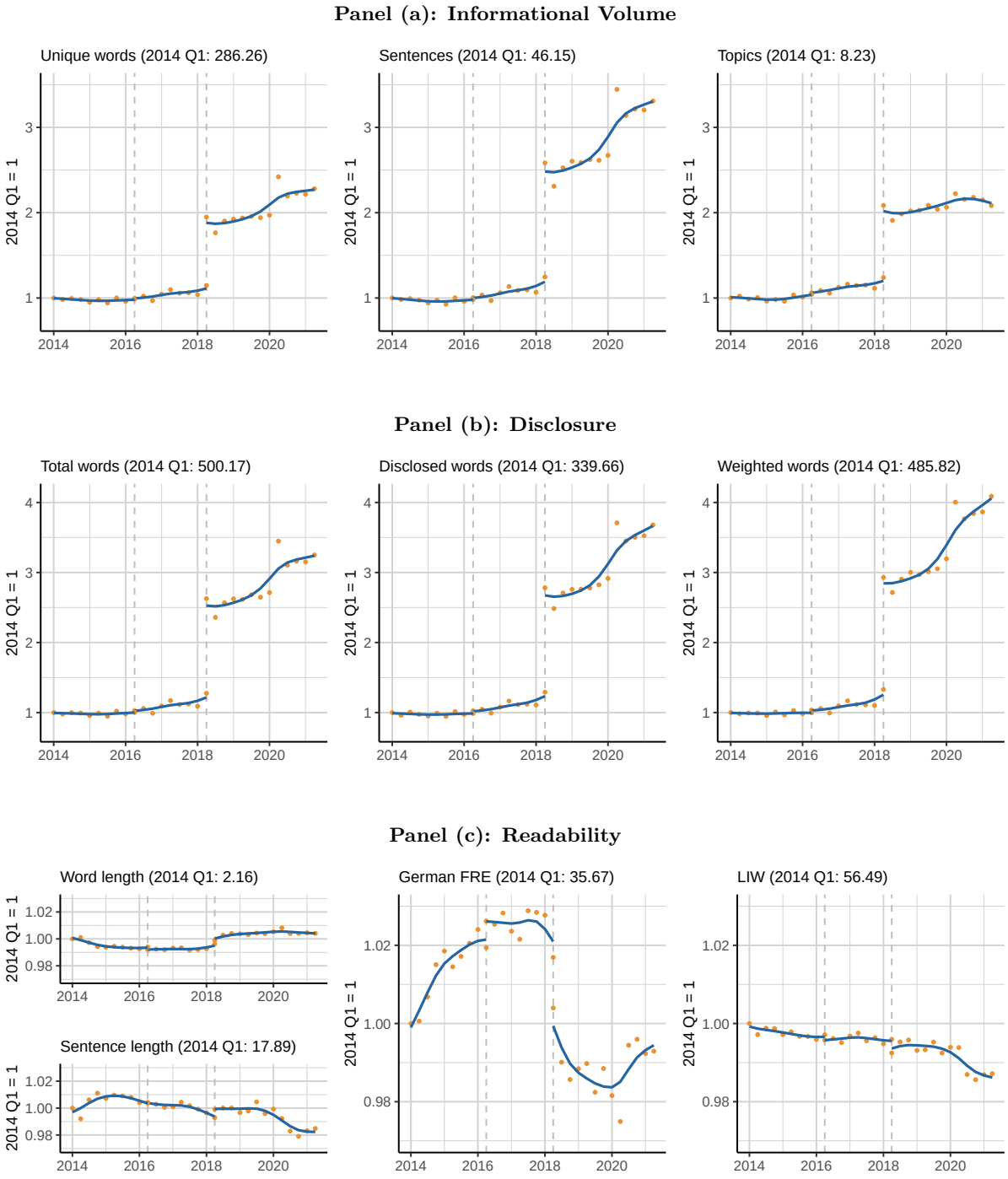
[43] In Figure A.8, we juxtapose the performance of readability scores (measured in the increase in node purity) and their popularity (measured in the number of Google Scholar citations). The figure illustrates that popularity in the literature and performance are not positively correlated.

[44] This difference in the LIW score implies a 15–20 percentage point decrease in the alignment of human and score-based readability of texts. See Figure A.6 and the figure notes.

[45] We plot total words in panel (b) to ease comparison. For levels of all measures, see Figure A.9 in the Appendix.

[46] We see no such effect of the passage of the GDPR in Q2 2016 (i.e., no announcement effect) and only a relatively small increase of about 25% (across all measures) leading up to Q2 2018.

Figure 4: Volume, Disclosure, and Readability



Notes: This figure presents quarterly averages of policy-level measures for informational volume (panel (a)), disclosure (panel (b)), and readability (panel (c)). Dots represent quarterly averages (values are normalized, with Q1 2014 = 1), the curves are fitted to the data (spline). The vertical dashed lines indicate the GDPR passage in Q2 2016 and GDPR enforcement in Q2 2018. Figure A.9 presents the non-normalized quarterly averages.

Table 5: Baseline Table

Panel (a): Informational Volume			
Dependent variable (in log):	**Unique words**	**Sentences**	**Topics**
	(1)	(2)	(3)
Post GDPR (=1)	0.5182***	0.7219***	0.4886***
	(0.0048)	(0.0064)	(0.0074)
Firm FE, Year FE	Yes	Yes	Yes
# Firm FE	64,605	64,605	64,596
R^2	0.789	0.793	0.696
Observations	409,221	409,221	409,071

Panel (b): Disclosure			
Dependent variable (in log):	**Total words**	**Disclosed words**	**Weighted words**
	(4)	(5)	(6)
Post GDPR (=1)	0.7078***	0.7929***	0.7775***
	(0.0065)	(0.0084)	(0.0073)
Firm FE, Year FE	Yes	Yes	Yes
# Firm FE	64,605	64,605	64,605
R^2	0.792	0.779	0.782
Observations	409,221	409,221	409,221

Notes: We report the results of fixed-effects OLS regressions (firm FE and year FE). Dependent variables are measures of information volume (panel (a)) and disclosure (panel (b)). All dependent variables are in log. Additional control variables are log Employees (as measure of size) and HHI (as measure of market concentration). Clustered (firms) standard errors in parentheses. Signif. levels: ***: 0.01, **: 0.05, *: 0.1.

(and the amount of disclosed information) increases as firms continue to adapt to the new regulatory regime. A disproportionate amount of additional volume is related to disclosure (Art. 13 and 14 GDPR) as both disclosed words and weighted words exhibit a stronger increase (and trend) post-GDPR.

The time series in Figure 4 are simple unconditional means, not accounting for observed or unobserved heterogeneity. In Table 5, we present fixed-effects OLS regression results. Our main variable of interest is a dummy variable equal to one for all post-GDPR observations, and zero otherwise. All dependent variables are in log, and we can interpret the post-GDPR as a percentage change of our dependent variable. We control for firm size (log employees) and market concentration (HHI). We also use firm fixed effects and year fixed effects to capture unobserved heterogeneity across firms and time. The results align well with our descriptive evidence in Figure 4. Policies in the post-GDPR period are on average 50–70% longer than in the pre-GDPR period. Moreover, they disclose almost 80% more content. All estimation coefficients are statistically significant at the 1% level.

Following the line of reasoning in Johnson et al. (forthcoming) or Peukert et al. (2022) (for a discussion, see Johnson (forthcoming)), we attribute the sudden change in our

volume and disclosure outcome variables (depicted in the top panels of Figure 4) to the GDPR-induced change in regulatory stringency itself.

The results in this section suggest that firms redrafted their privacy policies to comply with the new rules, particularly those laid out in Articles 13 and 14 GDPR that requested additional information. The obligation requiring firms to inform users about the processing of data, of course, is not an entirely new concept in the EU legal order. Prior to the GDPR, the Data Protection Directive (DPD) already required firms to inform data subjects about the identity of the data controller as well as the purposes of the processing of the data (Art. 10 DPD). With the entry into force of the GDPR, however, legal requirements of privacy policies have been fundamentally transformed. The GDPR introduces new categories a data subject has to be informed about. Examples include the legal basis for the processing of the data (Art. 13(1)(c) GDPR) and information about the rights for data subjects, such as the right to rectification, data portability, or the erasing of personal data. The changes we observe therefore capture additional information privacy policies now contain.

7.2 Readability Before and After the GDPR

Panel (c) in Figure 4 depicts the (normalized) quarterly averages of word length and sentence length as well as the readability scores German FRE and LIW. Note that the scale of the vertical axis is smaller than in the two top panels of the figure. A change from 1 to 1.01 translates to a 1% increase of the respective variable.

The effect of the GDPR on readability is ambiguous. Average word length is increasing (implying more difficult to read policies), whereas average sentence length decreases after an initial increase (and continues to follow what seems to be a general negative trend), suggesting an improvement of readability. We observe a similar ambiguity of results for our readability scores. The German FRE decreases by about 3% post-GDPR, implying a decrease in readability of privacy policies. The decrease of the LIW, albeit weak (but more pronounced in later periods) means an increase in readability.

In Table 6, we present fixed-effects regression results. We use firm fixed effects and year fixed effects in all models, but we include firm size (log employees) and market concentration (HHI) only in the models in panel (b). The results do not change with the inclusion of firm-level and industry-level characteristics. We find that post-GDPR policies use longer words (by 0.8%) but shorter sentences (by 1%). Both readability scores decrease after Q2 2018. The German FRE is 4% lower for post-GDPR policies, implying a decline in readability. The LIW is 0.4% lower for post-GDPR policies, implying an improvement of readability. Both effects are statistically significant at the 1% level.

Prediction 1 states that, if any, the effect of more regulatory stringency (post GDPR) is weak. This may be an explanation for the small and ambiguous results in Table 6. An

Table 6: Readability Factors and Scores

Panel (a): Baseline results (without firm-level controls)				
Dependent variable (in log):	**Word length**	**Sentence length**	**German FRE**	**LIW**
	(1)	(2)	(3)	(4)
Post GDPR (=1)	0.0079***	-0.0102***	-0.0392***	-0.0039***
	(0.0002)	(0.0012)	(0.0014)	(0.0005)
Firm FE, Year FE	Yes	Yes	Yes	Yes
# Firm FE	75,683	75,683	75,680	75,683
R^2	0.643	0.605	0.592	0.612
Observations	585,329	585,329	585,145	585,329

Panel (b): Full results				
Dependent variable (in log):	**Word length**	**Sentence length**	**German FRE**	**LIW**
	(5)	(6)	(7)	(8)
Post GDPR (=1)	0.0081***	-0.0102***	-0.0418***	-0.0041***
	(0.0003)	(0.0014)	(0.0018)	(0.0006)
log Employees	0.0007***	0.0011	-0.0030**	0.0004
	(0.0003)	(0.0014)	(0.0015)	(0.0006)
Concentration (HHI in '00)	-0.00001**	0.0000002	0.0001***	-0.00002*
	(0.000007)	(0.00003)	(0.00004)	(0.00001)
Firm FE, Year FE	Yes	Yes	Yes	Yes
# Firm FE	64,605	64,605	64,602	64,605
R^2	0.678	0.642	0.624	0.648
Observations	409,221	409,221	409,131	409,221

Notes: We report the results of fixed-effects OLS regressions (firm FE and year FE). Dependent variables are measures of readability (average word length, average sentence length, German FRE, and LIW). All dependent variables are in log. The results in panel (a) are without firm-level and industry-level characteristics log Employees (as measure of size) and HHI (as measure of market concentration). Clustered (firms) standard errors in parentheses. Signif. levels: ***: 0.01, **: 0.05, *: 0.1.

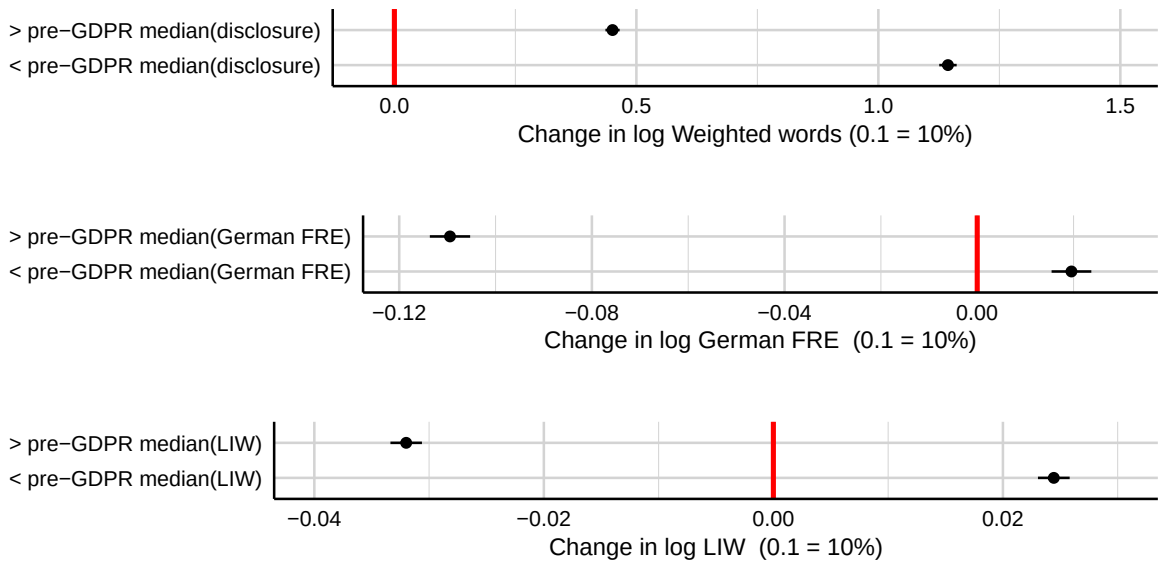
alternative explanation is the difficulty of measuring readability (as is our underlying presumption), with different scores or indices focusing on different textual components. We see that average word length and average sentence length (both expected to imply lower readability) respond differently to the introduction of the GDPR, and so will readability scores that assign different weights.

We provide supportive evidence of the measurement issues in Figure A.10, where we plot the GDPR coefficient for 35 readability scores. The results for 16 indices imply an improvement in readability post-GDPR, while the results for 19 indices imply a decline of readability. We interpret these results as a confirmation of our initial assumption: readability is difficult to measure, and enforcing this aspect of the transparency requirement is difficult.[47]

Panel (b) in Table 6 also reports the coefficients for firm size and market concentration. We find that policies in more concentrated markets are more readable (higher German

[47]Note that these findings also pose a challenge for researchers using readability indices indiscriminately, without considering the nature of the text domain.

Figure 5: Effect of GDPR Exposure ("Treatment Intensity") on Readability



Notes: This figure depicts the GDPR-associated change in disclosure (panel (a)) and readability (panels (b) and (c)) conditional on pre-GDPR levels of disclosure and readability. We include firm-level and industry-level characteristics log Employees (as measure of size) and HHI (as measure of market concentration). We report the GDPR coefficients (and 95% confidence intervals) for firms that had pre-GDPR policies above and below the median of pre-GDPR means of weighted words, German Flesch, and LIW (i.e., exposure). "< median(German FRE)" implies less readable policies, "< median(LIW)" implies more readable policies.

FRE, lower LIW) whereas larger firms have less readable privacy policies (lower German FRE). We further explore the firm and industry-level heterogeneity of the GDPR effect in Section 8.

7.3 GDPR Exposure as Treatment Intensity

In Figure 5, we present disclosure and readability results for firms with different levels of exposure to the GDPR (see, e.g. Chen et al., 2022) or varying "treatment intensity." We expect policies with high levels of disclosure pre-GDPR to exhibit smaller changes around the enforcement of the GDPR in Q2 2018 than short policies with little disclosure. Likewise, firms with policies that exhibit high readability pre-GDPR will adjust less to the GDPR's readability requirement than firms with difficult to read policies.

For our exposure results, we re-estimate model (6) in Table 5 (weighted words) and models (7) and (8) in Table 6 (German FRE and LIW) and interact the post-GDPR dummy with a dummy indicating if a firm had a pre-GDPR policy with weighted words, German FRE, and LIW above or below the median of the respective dependent variable. Figure 5 plots the conditional effects and 95% confidence intervals.

We find that higher treatment intensity (or higher GDPR exposure) triggers stronger

effects. Firms with below-median disclosure prior to Q2 2018 add significantly disclosure content to their privacy policies. Moreover, policies by firms with below-median readability prior to Q2 2018 exhibit a significant increase in readability, whereas the readability of above-median policies (policies that were already highly readable) declined.[48]

7.4 Hypothetical Compliance Rates

We conclude this section with a data exercise to derive hypothetical compliance rates. In the United States, state-level regulation of insurance contracts requires a minimum level of the Flesch Reading Ease score. Because of a lack of tangible regulatory guidelines for privacy policies, suppose one of the thresholds used in the U.S. (40 for insurance contracts in Massachusetts and Michigan, 45 in Florida, and 50 in Texas) is the respective threshold for readability of privacy policies. Given such a threshold, we can calculate compliance rates over time.

Because our policies are in German, we first determine the German equivalents of the U.S. thresholds (for English-language insurance contracts). We use a sample of 118,356 English-language privacy policies (posted pre-GDPR, between 2014 and 2017) compiled by Amos et al. (2021), calculate their respective FRE score, and determine the percentages of policies with scores below 40, 45, and 50. Assuming that the distributions for German and English FRE scores are the same but for a shift parameter, we can then calculate the percentiles of these percentage thresholds: 39.58, 41.83, and 44.85. We use these values as the German-language equivalents of the regulatory thresholds 40, 45, and 50.

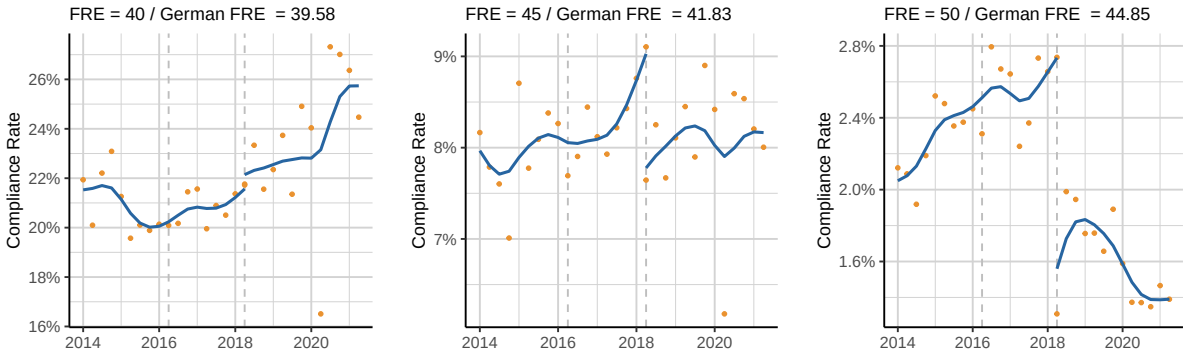
In Figure 6, we show the quarterly compliance rates for each of the three thresholds, in order of increasing stringency. The hypothetical compliance rates for the least stringent threshold increase in response to the GDPR, whereas hypothetical compliance decreases for the most stringent threshold (that requires higher readability).

8 The Role of Regulatory Scrutiny and Capacity

We now explore the role of regulatory scrutiny (Prediction 2) and capacity (Prediction 3). Prediction 2 states that firms under higher regulatory scrutiny will respond to the GDPR with more readable policies. The situation is similar for larger firms. Prediction 3 states that firms who are more likely to face an unconstrained regulator (with higher regulatory capacity) exhibit better compliance with the readability requirement. The effect on disclosure, if any, is likely much weaker (because firms comply with the disclosure requirement regardless of the regulator's capacity).

[48] For the German FRE, above pre-GDPR median(German FRE) implies high pre-GDPR readability. The GDPR effect on readability is negative (a decline in readability). For the LIW, above pre-GDPR median(LIW) implies low pre-GDPR readability. The GDPR effect on readability is negative (an increase in readability).

Figure 6: Hypothetical Compliance Rates



Notes: This figure depicts the compliance rate for the German equivalents of regulatory thresholds for the Flesch Reading Ease score of 40 (insurance contracts in Massachusetts and Michigan), 45 (Florida), and 50 (Texas). We determine these German equivalent thresholds as 39.58, 41.83, and 44.85. We define the compliance rate as the quarterly % of all policies with German FRE at or above the respective critical thresholds.

Table 7: Effect of Firm Size on Compliance

	Disclosure	**Readability**			
Dependent variable (in log):	**Weighted words**	**Word length**	**Sentence length**	**German FRE**	**LIW**
Post GDPR (=1)	0.7269***	0.0090***	-0.0136***	-0.0454***	-0.0037***
	(0.0101)	(0.0004)	(0.0021)	(0.0024)	(0.0008)
log Employees	0.0033	0.0010***	0.00005	-0.0042***	0.0005
	(0.0066)	(0.0003)	(0.0015)	(0.0016)	(0.0006)
Post GDPR (=1) × log Employees	0.0216***	-0.0004***	0.0015**	0.0017**	-0.0002
	(0.0032)	(0.0001)	(0.0007)	(0.0008)	(0.0003)
Firm FE, Year FE	Yes	Yes	Yes	Yes	Yes
# Firm FE	65,863	65,863	65,863	65,859	65,863
R^2	0.784	0.680	0.645	0.627	0.650
Observations	413,249	413,249	413,249	413,154	413,249

Notes: We report the results of fixed-effects OLS regressions (firm FE and year FE). Dependent variables are measures of disclosure (weighted words) and readability (average word length, average sentence length, German FRE, and LIW). All dependent variables are in log. Clustered (firms) standard errors in parentheses. Signif. levels: ***: 0.01, **: 0.05, *: 0.1.

8.1 Scrutiny: Firm Size and Market Concentration

We propose two proxies for regulatory scrutiny. We expect large firms (in log employees) and firms in concentrated markets (in HHI) to be subject to higher regulatory scrutiny. We re-estimate model (6) in Table 5 (weighted words) and models (5) through (8) in Table 6 (word length, sentence length, German FRE, and LIW) and interact the post-GDPR dummy with log employees (Table 7) and market-level HHI (Table 8).

The results in Table 7 highlight the role of firm size. First, large firms increase their compliance with the disclosure requirement in response to the GDPR (the interaction term for weighted words is positive and statistically significant). Moreover, larger firms

Table 8: Effect of Market Concentration on Compliance

Dependent variable (in log):	**Disclosure**	**Readability**			
	Weighted words	**Word length**	**Sentence length**	**German FRE**	**LIW**
Post GDPR (=1)	0.7509***	0.0081***	-0.0113***	-0.0401***	-0.0039***
	(0.0062)	(0.0002)	(0.0012)	(0.0015)	(0.0005)
Concentration (HHI in '00)	0.0006***	0.00002**	-0.0002***	0.000002	-0.00003
	(0.0002)	(0.000008)	(0.00004)	(0.00005)	(0.00002)
Post GDPR (=1) ×	0.00001	-0.00005***	0.0003***	0.0002**	0.00002
Concentration (HHI in '00)	(0.0003)	(0.00001)	(0.00006)	(0.00007)	(0.00002)
Firm FE, Year FE	Yes	Yes	Yes	Yes	Yes
# Firm FE	73,493	73,493	73,493	73,490	73,493
R^2	0.757	0.643	0.606	0.592	0.612
Observations	567,166	567,166	567,166	566,997	567,166

Notes: We report the results of fixed-effects OLS regressions (firm FE and year FE). Dependent variables are measures of disclosure (weighted words) and readability (average word length, average sentence length, German FRE, and LIW). All dependent variables are in log. Clustered (firms) standard errors in parentheses. Signif. levels: ***: 0.01, **: 0.05, *: 0.1.

increase word length less and sentence length more than smaller firms. Last, larger firms improve readability of their policies relative to small firms. Average German FRE readability declines post-GDPR, but this effect is weaker for larger firms. In line with Prediction 2, their non-compliance is less severe than that of smaller firms. We do not find any statistically significant results for the LIW.

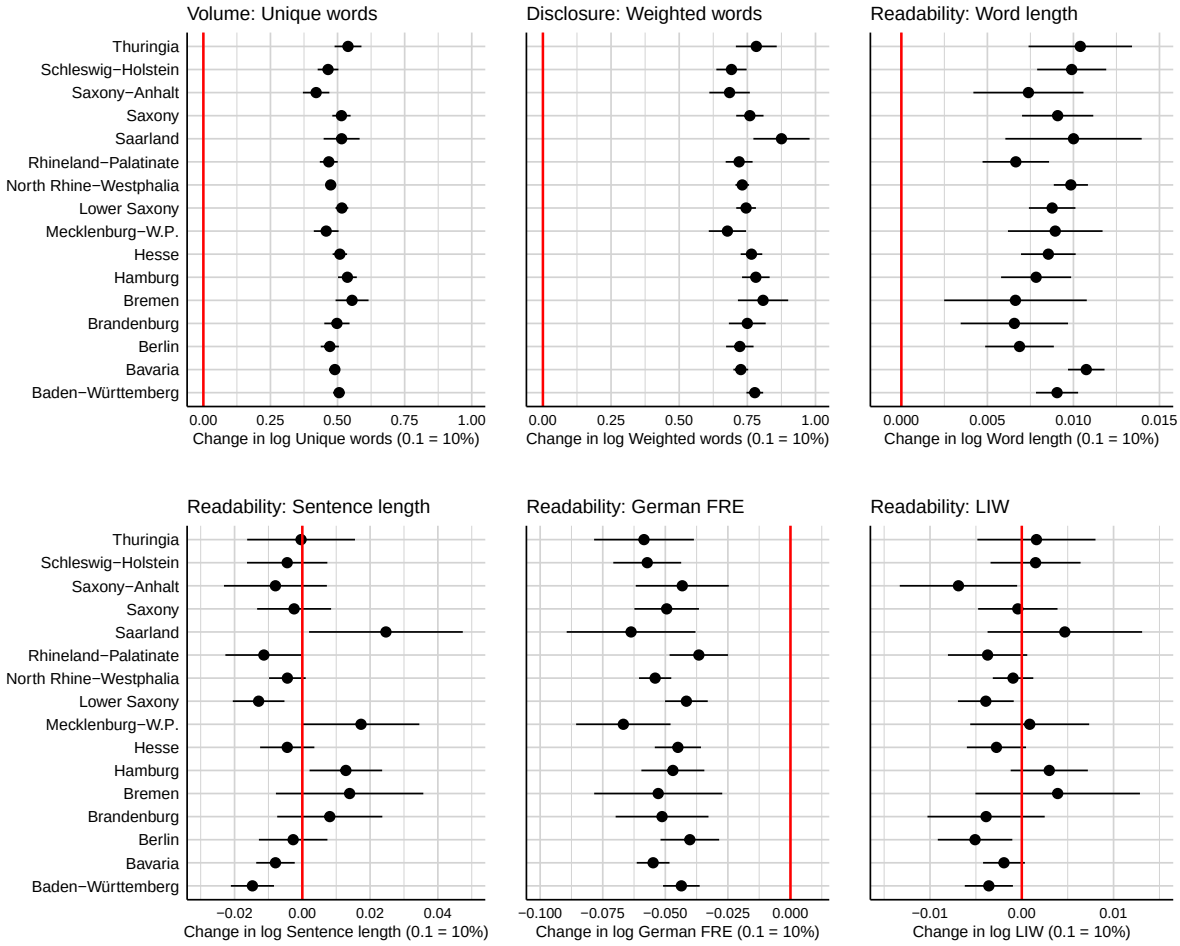
Table 8 summarizes our results for the effect of market concentration on compliance. First, we find no effect of market concentration on disclosure compliance: the interaction term is insignificant. Second, firms in more concentrated industries increase word length less but sentence length more than other firms. And last, firms in more concentrated industries reduce readability of their policies less than other firms. This latter result is in line with Prediction 2.

Our results for firm size and market concentration fall well in line with our predictions, except for the effects on disclosure compliance. This divergence of results is possibly driven by how larger firms are also expected to have lower drafting costs (reducing the costs of disclosure compliance) and may simply have more information to disclose (as their business may also be broader than that of small firms). Especially that latter factor should not have any bearing on the results for readability compliance.

8.2 Capacity: State-Level Regulators

We use budget numbers for 16 state DPAs to measure the resources and capacity of state regulators, exploiting considerable variation across states (see Figure 3) and over time (see Figures A.1, A.2, and A.3). The underlying assumption is that DPAs with larger budgets are less likely constrained, and the theoretical implications from our model summarized

Figure 7: State-Level Results (Volume, Disclosure, and Readability)



Notes: This figure depicts the GDPR-associated change in volume (unique words), disclosure (weighted words), and readability (average word length, average sentence length, German FRE, and LIW) conditional on the firm's home state. We report the GDPR coefficients (and 95% confidence intervals). All models control for log Employees (to measure firm size) and HHI (to measure for market concentration).

in Prediction 3 apply.

We show a set of preliminary results in Figure 7 that presents the conditional effects of the GDPR on measures of volume, disclosure, and readability at the state level. We see little heterogeneity for volume (unique words) and disclosure (weighted words). Also, the increase of average word length after the GDPR is relatively uniform across states, whereas the effect on sentence length exhibits considerable heterogeneity. Last, we observe similar patterns for our readability scores. The GDPR effect on German FRE is consistently negative (reducing readability) wheres the effect on LIW is inconsistent across states.

We re-estimate model (6) in Table 5 (weighted words) and models (5) through (8) in Table 6 (word length, sentence length, German FRE, and LIW) and interact the post-GDPR dummy with one of three budget variables: total budget (per capita), staff budget (per capita), and overall budget share. All three budget variables are lagged. We present

Table 9: Effect of Regulatory Capacity on Compliance

Dependent variable (in log):	**Disclosure**	**Readability**			
	Weighted words	**Word length**	**Sentence length**	**German FRE**	**LIW**
Panel (a): DPA Budget – Total Budget Per Capita					
× Total budget (per capita, lagged)	-0.0087	-0.0022***	0.0087***	0.0048*	0.0008
	(0.0117)	(0.0005)	(0.0024)	(0.0029)	(0.0010)
# Firm FE	74,576	74,576	74,576	74,573	74,576
R^2	0.757	0.643	0.605	0.591	0.611
Observations	579,132	579,132	579,132	578,955	579,132
Panel (b): DPA Budget – Personnel Budget Per Capita					
× Staff budget (per capita, lagged)	-0.0165	-0.0027***	0.0109***	0.0053	0.0005
	(0.0138)	(0.0006)	(0.0029)	(0.0034)	(0.0012)
# Firm FE	74,576	74,576	74,576	74,573	74,576
R^2	0.757	0.643	0.605	0.591	0.611
Observations	579,132	579,132	579,132	578,955	579,132
Panel (c): DPA Budget – Share of Overall State Budget					
× Overall share (lagged)	-78.98	-10.70***	68.50***	5.501	9.485*
	(65.37)	(2.552)	(14.12)	(16.24)	(5.714)
# Firm FE	74,576	74,576	74,576	74,573	74,576
R^2	0.757	0.643	0.605	0.591	0.611
Observations	579,132	579,132	579,132	578,955	579,132

Notes: We report the results of fixed-effects OLS regressions (firm FE and year FE). Dependent variables are measures of disclosure (weighted words) and readability (average word length, average sentence length, German FRE, and LIW). We report the interaction term of the Post GDPR (=1) dummy and a budget variable (budgeted total expenditure per capita in panel (a); budgeted personnel expenditure per capita in panel (b); total DPA budget over state's total expenditure in panel (c)). All dependent variables are in log. Clustered (firms) standard errors in parentheses. Signif. levels: ***: 0.01, **: 0.05, *: 0.1.

the results in Table 9.

First, we do not see an effect of regulatory capacity on firms' disclosure compliance. In fact, all point estimates are negative and, if anything, hint at weaker compliance with the disclosure requirement in states with higher-budget regulators.[49] Our model predicts negative effects on disclosure compliance for firms with sufficiently high compliance costs. Second, we see mixed results for the readability requirement. Firms with higher-budget regulators increase the average word length of their privacy policies less than other firms. At the same time, the average sentence length decreases by less than that of other firms. The results for the German FRE are well in line with our prediction. Firms in higher-budget states exhibit better readability compliance than other firms – all interaction terms are positive. The coefficient for total budget is significant at the 10%; the coefficient for staff budget has a p-value of 0.12. We do not see these patterns for LIW. On the contrary,

[49]The statistically strongest coefficient (staff budget) has a p-value of 0.23.

a higher overall budget share induces less readability compliance.

Overall, the effects of state regulators' budgets on firms' compliance comports fairly well with our theoretical predictions. Because firms' disclosure compliance is at high baseline level (see Table 5), additional regulatory capacity has little effect on firm behavior. For readability, firms facing a higher-budget regulator anticipate stronger enforcement of the readability requirement. We see some evidence of improved readability compliance in response to stronger regulatory capacity.

9 Conclusion

We study firms' compliance with the transparency requirement of the GDPR, compelling firms to disclose information about the nature of their data collection, processing, and use in a "concise, transparent, intelligible and easily accessible form, using clear and plain language" (Art. 12(1) GDPR). Disclosure is objective and compliance verifiable. Readability, on the other hand, is subjective and vague, rendering it difficult to enforce. We show in a simple theoretical framework that this asymmetry in enforceability will lead to differential dynamics in firms' compliance. Firms will anticipate regulators to enforce what is indeed enforceable, and then comply accordingly.

We take these theoretical insights to the data, using a sample of some 585,000 privacy policies posted by more than 75,000 German firms between 2014 and 2021. We find strong evidence for disclosure compliance but weak evidence for readability compliance. Assuming that larger firms and those in more concentrated markets are primary targets for data protection authorities, we use firm size and 4-digit industry Herfindahl-Hirschman Indices to proxy for regulatory scrutiny. We find that better readability compliance for both, and better disclosure compliance only for firm size. The relative patterns are in line with your theoretical predictions: firms already exhibit high disclosure compliance, more regulatory scrutiny should not have a meaningful effect on their disclosure (relative to readability).

Last, our data is set in the unique regulatory landscape in Germany, with 16 data protection authorities tasked with enforcing EU data protection law. Each of these authorities has its own budget. We exploit across-state and time variation in the authorities' budgets to examine the effect of a regulator's budget constraint (and the impact that has on its enforcement activities) on the respective firms' compliance. Our data confirms that a regulator's constraint does not affect firms' disclosure compliance. However, we find some evidence that firms in states with higher-budget data protection authorities (that can be expected to engage more actively in the enforcement of the readability requirement) exhibit better readability compliance.

Our results have immediate implications for enforcement activities of data protection authorities and can explain the falling short of the GDPR of its potential (European

Commission, 2019). Our results also speak more generally to the effectiveness (or the lack thereof) of regulatory tools that are based on unverifiable (or difficult to verify) information.

References

Amos, Ryan, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer (2021) "Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset," in *Proceedings of The Web Conference 2021*, WWW '21, 22: Association for Computing Machinery.

Amstad, Toni (1978) *Wie verständlich sind unsere Zeitungen?* Ph.D. dissertation, University of Zurich.

Armstrong, Mark and David E. M. Sappington (2006) "Regulation, Competition, and Liberalization," *Journal of Economic Literature*, 44 (2), 325–366.

Art. 29 Working Party (2018) "Article 29 Working Party: Guidelines on Transparency Under Regulation 2016/679 (wp260rev.01)," The Working Party on the Protection of Individuals with Regard to the Processing of Personal Data, available for download at https://ec.europa.eu/newsroom/article29/items/622227.

Bakos, Yannis, Florencia Marotta-Wurgler, and David R. Trossen (2014) "Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts," *Journal of Legal Studies*, 43 (1), 1–35.

Ball, Christopher, Gerard Hoberg, and Vojislav Maksimovic (2015) "Disclosure, Business Change and Earnings Quality," unpublished manuscript, available at http://papers.ssrn.com/abstract=2260371.

Barbaresi, Adrien (2019) "German Political Speeches Corpus (Version v4.2019) [Data set]," Zenodo. https://doi.org/10.5281/zenodo.3611246.

Bardsley, Peter (1996) "Tax Compliance Games with Imperfect Auditing," *Public Finance*, 51, 473–489.

Becher, Shmuel and Uri Benoliel (2021) "Law in Books and Law in Action: The Readability of Privacy Policies and GDPR," in Mathis, Klaus and Avishalom Tor eds. *Consumer Law and Economics*, 179–204, Cham, Switzerland: Springer.

Bellstam, Gustaf, Sanjai Bhagat, and J. Anthony Cookson (2021) "A Text-Based Analysis of Corporate Innovation," *Management Science*, 67 (7), 4004–4031.

Benoit, Kenneth, Kevin Munger, and Arthur Spirling (2019) "Measuring and Explaining Political Sophistication Through Textual Complexity," *American Journal of Political Science*, 63 (2), 491–508.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo (2018) "quanteda: An R Package for the Quantitative Analysis of Textual Data," *Journal of Open Source Software*, 3 (30), 774–777.

Bersch, Johannes, Sandra Gottschalk, Bettina Müller, and Michaela Niefert (2014) "The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany," ZEW Discussion Paper 14-104, ZEW – Leibniz Centre for European Economic Research, Mannheim, Germany.

Björnson, Carl-Hugo (1968) "Läsbarhet [Readability]," *Stockholm: Liber*.

Blei, David M. (2012) "Probabilistic Topic Models," *Communications of the ACM*, 55 (4), 77–84.

Blei, David M. and John D. Lafferty (2009) "Topic Models," in Srivastava, Ashok N. and Mehran Sahami eds. *Text Mining: Classification, Clustering, and Applications*, Boca Raton, FL: CRC Press.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003) "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Bradley, Ralph A. and Milton E. Terry (1952) "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, 39 (3/4), 324–345.

Brody, Samuel and Noemie Elhadad (2010) "An Unsupervised Aspect-Sentiment Model for Online Reviews," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 804–812, Association for Computational Linguistics.

Chen, Chinchih, Carl Benedikt Frey, and Giorgio Presidente (2022) "Privacy Regulation and Firm Performance: Estimating the GDPR Effect Globally," The Oxford Martin Working Paper Series on Technological and Economic Change Working Paper No. 2022-1, Oxford Martin School, Oxford, UK.

Courtis, John K. (1995) "Readability of Annual Reports: Western versus Asian Evidence," *Accounting, Auditing and Accountability Journal*, 8 (2), 4–17.

DataGrail (2019) "The Age of Privacy: The Cost of Continuous Compliance. Benchmarking the Ongoing Operational Impact of GDPR & CCPA," Technical report, available at https://www.datagrail.io/resources/reports/gdpr-ccpa-cost-report/.

Degeling, Martin, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz (2019) "We Value Your Privacy … Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy," *Proceedings of the 26th Network and Distributed System Security Symposium (NDSS'19)*.

European Commission (2019) "Data Protection Rules as a Trust-Enabler in the EU and Beyond – Taking Stock," Communication from the Commission to the European Parliament and the Council, European Commission.

European Data Protection Board (2020) "Contribution of the EDPB to the Evaluation of the GDPR under Article 97," Technical report, European Data Protection Board.

Ezat, Amr N. (2019) "The Impact of Earnings Quality on the Association Between Readability and Cost of Capital: Evidence from Egypt," *Journal of Accounting in Emerging Economies*, 9 (3), 366–385.

Fellingham, John C. and Paul Newman (1985) "Strategic Considerations in Auditing," *Accounting Review*, 60 (1), 634–650.

Flesch, R. (1948) "A New Readability Yardstick," *Journal of Applied Psychology*, 32 (3), 221–233.

Frankenreiter, Jens (2022) "Cost-Based California Effects," *Yale Journal on Regulation*, 39, 1155–1217.

Ganglmair, Bernhard and Malcolm I. Wardlaw (2017) "Complexity, Standardization, and the Design of Loan Agreements," Unpublished manuscript, University of Georgia, available at https://ssrn.com/abstract=2952567.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019) "Text as Data," *Journal of Economic Literature*, 57 (3), 535–574.

Goldberg, Samuel, Garrett Johnson, and Scott Shriver (2019) "Regulating Privacy Online: The Early Impact of the GDPR on European Web Traffic & E-Commerce Outcomes," Unpublished manuscript, available at https://ssrn.com/abstract=3421731.

Graetz, Michael, Jennifer Reinganum, and Louis Wilde (1986) "The Tax Compliance Game: Toward an Interactive Theory of Law Enforcement," *Journal of Law, Economics, and Organization*, 2 (1), 1–32.

Greenberg, Joseph (1984) "Avoiding Tax Avoidance: A (Repeated) Game-Theoretic Approach," *Journal of Economic Theory*, 32 (1), 1–13.

Griffiths, Thomas L. and Mark Steyvers (2004) "Finding Scientific Topics," *Proceedings of the National Academy of Sciences*, 101 (suppl 1), 5228–5245, 10.1073/pnas.0307752101.

Grün, Bettina and Kurt Hornik (2011) "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software*, 40 (13), 1–30.

Gustafsson, Marita (1984) "The Syntactic Features of Binomial Expressions in Legal English," *Text: An Interdisciplinary Journal for the Study of Discourse*, 4 (1-3), 123–142.

Hall, David, Daniel Jurafsky, and Christopher D. Manning (2008) "Studying the History of Ideas Using Topic Models," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 363–371.

Heyes, Anthony G. (1994) "Environmental Enforcement when 'Inspectability' is Endogenous: A Model with Overshooting Properties," *Environmental and Resource Economics*, 4, 479–494.

Jensen, Carlos and Colin Potts (2004) "Privacy Policies as Decision-Making Tools," in Dykstra-Erickson, Elizabeth and Manfred Tscheligi eds. *Proceedings of the 2004 Conference on Human Factors in Computing Systems (CHI '04)*, 471–478, New York, N.Y.: ACM.

Johnson, Garrett (forthcoming) "Economic Research on Privacy Regulation: Lessons from the GDPR and Beyond," in Goldfarb, Avi and Catherine Tucker eds. *The Economics of Privacy*, Chicago, Ill.: University of Chicago Press.

Johnson, Garrett, Scott Shriver, and Samuel Goldberg (forthcoming) "Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR," *Management Science*.

Kellermann, Gudrun (2014) "Leichte und Einfache Sprache – Versuch einer Definition," *Aus Politik und Zeitgeschichte*, 64 (9-11), 7–10.

Kinne, Jan and Janna Axenbeck (2019) "Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany," ZEW Discussion Paper 18-033, ZEW – Leibniz Centre for European Economic Research, Mannheim.

Koski, Heli and Nelli Valmari (2020) "Short-Term Impacts of the GDPR on Firm Performance," ETLA Working Papers 77, ETLA Economic Research.

Laffont, Jean-Jacques (1994) "The New Economics of Regulation Ten Years After," *Econometrica*, 62 (3), 507–537.

——— (2005) *Regulation and Development*, Cambrdge, UK: Cambridge University Press.

Larsen, Vegard H. and Leif A. Thorsrud (2019) "The Value of News for Economic Developments," *Journal of Econometrics*, 210 (1), 203–218.

Lin, Nick and Moritz Osnabrügge (2018) "Making Comprehensible Speeches When Your Constituents Need It," *Research and Politics*, 5 (3), 1–8.

Linden, Thomas, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz (2020) "The Privacy Policy Landscape After the GDPR," *Proceedings on Privacy Enhancing Technologies*, 2020 (1), 47–64.

Livermore, Michael A., Allen Riddell, and Daniel Rockmore (2016) "Agenda Formation and the US Supreme Court: A Topic Model Approach," *Arizona Law Review*, 1 (2).

Loughran, Tim and Bill McDonald (2016) "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, 54 (4), 1187–1230.

Macho-Stadler, Ines and David Perez-Castrillo (2006) "Optimal Enforcement Policy and Firms' Emissions and Compliance with Environmental Taxes," *Journal of Environmental Economics and Management*, 51 (1), 110–131.

Marotta-Wurgler, Florencia (2008) "Competition and the Quality of Standard Form Contracts: The Case of Software License Agreements," *Journal of Empirical Legal Studies*, 5 (3), 447–475.

McCallum, Andrew, Xuerui Want, and Andrés Corrada-Emmanuel (2007) "Topic and Role Diversity in Social Networks with Experiments on Enron and Academic Email," *Journal of Artificial Intelligence Research*, 30 (1), 249–272.

Mellinkoff, David (1963) *The Language of the Law*, Boston, Mass.: Little, Brown & Co.

Milne, George R., Mary J. Culnan, and Henry Greene (2006) "A Longitudinal Assessment of Online Privacy Notice Readability," *Journal of Public Policy & Marketing*, 25 (2), 238–249.

Möllers, Christoph, Anna Shadrova, and Luisa Wendel (2021) "BVerfGE-Korpus (1.0) [Data set]," Zenodo. https://doi.org/10.5281/zenodo.4551408.

Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer (2022) "Regulatory Spillovers and Data Governance: Evidence from the GDPR," *Marketing Science*, 41 (4), 318–340.

Polčák, Radim (2020) "Article 12. Transparent Information, Communication and Modalities for the Exercise of the Rights of the Data sSubject," in Kuner, Christopher, Christopher Docksey, and Lee Bygrave eds. *The EU General Data Protection Regulation: A Commentary*, 398–412, Oxford, UK: Oxford University Press.

Rauh, Christian (2023) "Clear Messages to the European Public? The language of European Commission Press Releases 1985–2020," *Journal of European Integration*, 45 (4), 683–701.

Ruckman, Karen and Ian P. McCarthy (2017) "Why Do Some Patents Get Licensed While Others Do Not?," *Industrial and Corporate Change*, 26 (4), 667–688.

Schütz, Philip (2018) "Zum Leben zu wenig, zum Sterben zu viel? Die finanzielle und personelle Ausstattung deutscher Datenschutzbehörden im Vergleich," in Roßnagel, Alexander, Michael Friedewald, and Marit Hansen eds. *Die Fortentwicklung des Datenschutzes: Zwischen Systemgestaltung und Selbstregulierung*, 251–268, Wiesbaden: Springer Vieweg.

Solove, Daniel J. (2013) "Introduction: Privacy Self-Management and the Consent Dilemma," *Harvard Law Review*, 126 (7), 1880–1903.

Stefanowitsch, Anatol (2014) "Leichte Sprache, komplexe Wirklichkeit," *Aus Politik und Zeitgeschichte*, 64 (9-11), 11–18.

Stern, Jon (2000) "Electricity and Telecommunications Regulatory Institutions in Small and Developing Countries," *Utilities Policy*, 9 (3), 131–157.

Steyvers, Mark and Thomas L. Griffiths (2011) "Probabilistic Topic Models," in Landauer, Thomas K., Danielle S. McNamara, Simon Dennis, and Walter Kintsch eds. *Handbook of Latent Semantic Analysis*, Chap. 21, 437–448, New York, NY: Routledge.

Taddy, Matthew A. (2012) "On Estimation for Topic Models," in Lawrence, Neil and Mark Girolami eds. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 22, 1184–1193.

Wagner, Isabel (2023) "Privacy Policies Across the Ages: Content of Privacy Policies 1996–2021," *ACM Transactions on Privacy and Security*, 26 (3), 1–32.

Waldman, Ari E. (2021) *Industry Unbound: The Inside Story of Privacy, Data, and Corporate Power*, Cambridge, UK: Cambridge University Press.

Wei, Xing and W. Bruce Croft (2006) "LDA-Based Document Models for Ad-Hoc Retrieval," in *SIGIR 2006 Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 178–185.

Wojahn, Oliver, Susanne Geister, and Julia Richter (2015) "The Impact of Analyst Report Complexity on Trading Decisions in an Experimental Setting," *Journal of Behavioral and Experimental Finance*, 7, 29–32.

Yuan, Bocong and Jiannan Li (2019) "The Policy Effect of the General Data Protection Regulation (GDPR) on the Digital Public Health Sector in the European Union: An Empirical Investigation," *International Journal of Environmental Research and Public Health*, 16 (1070), 1–15.

Zanfir-Fortuna, Gabriela (2020) "Article 13. Information to be Provided Where Personal Data are Collected from the Data Subject," in Kuner, Christopher, Christopher Docksey, and Lee Bygrave eds. *The EU General Data Protection Regulation: A Commentary*, 413–433, Oxford, UK: Oxford University Press.

A Formal Proofs

Proof of Proposition 1

The game is solvable by iterated dominance. First, $a_{d,r}$ is the regulator's (weakly) dominant strategy: given Assumption 2, the regulator strictly prefers to play $a_{d,r}$ if the firm chooses $(0,0)$, $(d,0)$, or $(0,r)$. Moreover, the regulator is indifferent (between all its actions) if the firms plays (d,r).

By iterated dominance, in equilibrium the firm must play a best response to the regulator's dominant strategy. It is sufficient to compare the payoffs of the firm when the regulator plays $a_{d,r}$. First, both $(0,0)$ and $(0,r)$ are dominated by $(d,0)$ in the reduced game. Strategy $(d,0)$ dominating $(0,r)$ follows once again from Assumption 2: $\pi_d > \pi_r$ implies

$$(1-\pi_r)(v-k) > (1-\pi_d)(v-k).$$

To see why $(d,0)$ dominates $(0,0)$, note that

$$(1-\pi_r)(v-k) > (1-\pi_d)(1-\pi_r)v \quad \Longleftrightarrow \quad k < \pi_d v.$$

The right-hand side holds by Assumptions 1 ($0 < k < \frac{v}{2}$) and 2 ($\pi_d > \pi_r > \frac{1}{2}$). Last, depending on the value of k, the firm chooses either $(d,0)$ or (d,r). It holds:

$$v - 2k > (1-\pi_r)(v-k) \quad \Longleftrightarrow \quad k < \frac{\pi_r}{1+\pi_r} v = k^u$$

Proof of Proposition 2

First, it is immediate to show that no pure strategy equilibria can exist when the regulator is constrained. The regulator's undominated strategies are a_d and a_r; none of the firm's strategies are dominated. Suppose the regulator played a_j with probability one. The firm best response would then be to play $(d,0)$ if $j = d$, or $(0,r)$ if $j = r$. Then, the regulator would want to deviate from his strategy. We look for mixed strategy equilibria: each agent plays their undominated strategies with some probability in a way that makes the other indifferent between their undominated strategies.

To make the regulator to be indifferent between a_d and a_r, the firm can play $(d,0)$ with probability p_d, $(0,r)$ with probability p_r, $(0,0)$ with probability $1-p_d-p_r$. Alternatively, she can play (d,r) with probability one. Notice that no mixed strategy involving (d,r) can exist, since the regulator's best response would be to optimally reply to the other strategy with probability one, which would make the firm want to deviate.

We find p_d, p_r that satisfy:

$$(1-p_d-p_r)(1-\pi_d)(-\gamma)-p_d\gamma-p_r(1-\pi_d)\gamma=(1-p_d-p_r)(1-\pi_r)(-\gamma)-p_d(1-\pi_r)\gamma-p_r\gamma$$

That is:

$$\begin{aligned} p_r &\in \left[0,\frac{\pi_r}{\pi_d+\pi_r}\right] \\ p_d &= \frac{\pi_d-(1-p_r)\pi_r}{\pi_d} \\ 1-p_d-p_r &= (1-p_r)\frac{\pi_r}{\pi_d}-p_r \end{aligned}$$

Since $p_d > 0$, the regulator must make the firm indifferent between $(d,0)$ and one between $(0,r)$ and $(0,0)$. Suppose that the regulator wanted to make the firm indifferent between $(d,0)$ and $(0,r)$; he must play a_d with probability $p^r_{a_d}$ that solves:

$$p^r_{a_d}(v-k)+\left(1-p^r_{a_d}\right)(1-\pi_r)(v-k)=p^r_{a_d}(1-\pi_d])(v-k)+\left(1-p^r_{a_d}\right)(v-k)$$

Suppose now that the regulator wanted to make the firm indifferent between $(d,0)$ and $(0,0)$ instead; then, the respective probability, $p^0_{a_d}$, that solves:

$$p^0_{a_d}(v-k)+\left(1-p^0_{a_d}\right)(1-\pi_r)(v-k)=p^0_{a_d}(1-\pi_d)v+\left(1-p^0_{a_d}\right)(1-\pi_r)v$$

which lead to probabilities:

$$\begin{aligned} p^r_{a_d} &= \frac{\pi_r}{\pi_d+\pi_r} \\ p^0_{a_d} &= \frac{(1-\pi_r)k}{\pi_d v-\pi_r k} \end{aligned}$$

We then have three candidate equilibria in which both players mix between two strategies; further, we must check when, if ever, the firms wants to deviate to (d,r). To do so, we obtain the utility of the firm when she mixes between $(d,0)$ and either $(0,r)$ or $(0,0)$ to determine which mixed equilibrium would emerge given parameters k, π_d, and π_r. Then, we compare the resulting utilities with the utility of full compliance, $v-2k$.

Suppose first that the regulator played $p^r_{a_d}=\frac{\pi_r}{\pi_d+\pi_r}$: we check for which parameters playing the firm does not want to deviate from the corresponding mixed strategy that would form an equilibrium, that is, mixing between $(d,0)$ and $(0,r)$ according to $p_d=\frac{\pi_d}{\pi_d+\pi_r}$. By plugging in $p^r_{a_d}$ in the expected utility of the firm under the various strategies,

we obtain:

$$\begin{aligned} E[(d,0)]|_{p^r_{a_d}} = E[(0,r)]|_{p^r_{a_d}} &= \frac{\pi_r}{\pi_d+\pi_r}(v-k) + \left(1-\frac{\pi_r}{\pi_d+\pi_r}\right)(1-\pi_r)(v-k) \\ &= \frac{(v-k)[\pi_d(1-\pi_r)+\pi_r]}{\pi_d+\pi_r} \end{aligned}$$

$$\begin{aligned} E[(0,0)]|_{p^r_{a_d}} &= \frac{\pi_r}{\pi_d+\pi_r}(1-\pi_d)\,v + \left(1-\frac{\pi_r}{\pi_d+\pi_r}\right)(1-\pi_r)\,v \\ &= \frac{v[\pi_d+\pi_r-2\pi_d\pi_r]}{\pi_d+\pi_r} \end{aligned}$$

Direct comparison reveals that, subject to the regulator mixing according to $p^r_{a_d}$, $E[(d,0)]|_{p^r_{a_d}} = E[(0,r)]|_{p^r_{a_d}} > E[(0,0)]|_{p^r_{a_d}}$ if and only if one of two conditions are satisfied:

$$\begin{gathered} \frac{1}{2} < \pi_r < \min\left(\pi_d, \frac{\pi_d}{3\pi_d-1}\right) \leq \frac{2}{3} \quad \wedge \quad k < \frac{\pi_r\pi_d}{\pi_r+\pi_d-\pi_r\pi_d}v, \\ \frac{\pi_d}{3\pi_d-1} < \pi_r < \pi_d < 1 \quad \wedge \quad k < \frac{v}{2}. \end{gathered}$$

Furthermore, $E[(d,0)]|_{p^r_{a_d}} > E[(d,r)]|_{p^r_{a_d}} = v-2k$ if and only if:

$$k < \frac{\pi_r\pi_d}{\pi_r+\pi_d+\pi_r\pi_d}v$$

Suppose now that the regulator played $p^0_{a_d} = \frac{(1-\pi_r)k}{\pi_d v-\pi_r k}$; we repeat the same exercise to check for which parameters an equilibrium in which the firm mixes between $(d,0)$ and $(0,0)$:

$$\begin{aligned} E[(d,0)]|_{p^0_{a_d}} = E[(0,0)]|_{p^0_{a_d}} &= \frac{(1-\pi_r)\,k}{\pi_d v-\pi_r k}(v-k) + \left(1-\frac{(1-\pi_r)\,k}{\pi_d v-\pi_r k}\right)(1-\pi_r)(v-k) \\ &= \frac{(v-k)v\pi_d(1-\pi_r)}{\pi_d v-\pi_r k} \end{aligned}$$

$$\begin{aligned} E[(0,r)]|_{p^0_{a_d}} &= \frac{(1-\pi_r)\,k}{\pi_d v-\pi_r k}(v-k) + \left(1-\frac{(1-\pi_r)\,k}{\pi_d v-\pi_r k}\right)(1-\pi_r)(v-k) \\ &= \frac{(v-k)[k\pi_d - v\pi_r + k\pi_r(1-\pi_d)]}{\pi_d v-\pi_r k} \end{aligned}$$

Again by direct comparison, it holds that $E[(d,0)]|_{p^0_{a_d}} = E[(0,0)]|_{p^0_{a_d}} > E[(0,r)]|_{p^0_{a_d}}$ iff:

$$\frac{1}{2} < \pi_r < \min\left(\pi_d, \frac{\pi_d}{3\pi_d-1}\right) \leq \frac{2}{3} \quad \wedge \quad k > \frac{\pi_r\pi_d}{\pi_r+\pi_d-\pi_r\pi_d}v$$

Combining the conditions above we immediately obtain the equilibria described in points 2, 3, and 4 of Proposition 2. It is clear that no other equilibria can exist for $k > \underline{k} = \frac{\pi_r \pi_d}{\pi_r + \pi_d + \pi_r \pi_d} v$ since no other deviations are available to the firm and no other undominated strategy is available to the regulator. Notice in particular that for $k < \underline{k}$ there cannot be any equilibrium in which the firm plays $(0,0)$ with positive probability from the above calculations. Therefore, the only comparison that matters for these parameters is a comparison between pure compliance, (d,r), and mixing between $(d,0)$, $(0,r)$.

From the calculations above, the former dominates the latter for $k < \underline{k} = \frac{\pi_r \pi_d}{\pi_r + \pi_d + \pi_r \pi_d} v$. This holds when the regulator mixes according to $p^r_{a_d}$. However, it can be shown that for $k < \underline{k}$ infinite payoff equilibria exist. In these equilibria, the firm plays (d,r) with probability one; the regulator mixes between a_d and a_r with different probabilities. To characterize them all, we find the highest and lowest probability of playing a_d as a function of k that makes the firm weakly better off playing (d,r) than deviating.

Recall that the expected utility of the firm playing $(d,0)$ and $(0,r)$ when the regulator plays a_d with probability $p^r_{a_d}$ are:

$$\begin{aligned} E[(d,0)]|_{p^r_{a_d}} &= p^r_{a_d}(v-k) + \left(1 - p^r_{a_d}\right)(1-\pi_r)(v-k) \\ E[(d,0)]|_{p^r_{a_d}} &= p^r_{a_d}(1-\pi_d)(v-k) + \left(1 - p^r_{a_d}\right)(v-k) \end{aligned}$$

We are interested in $p^r_{a_d}$ that makes the firm indifferent between $(d,0)$ and (d,r), and $p^r_{a_d}$ that makes the firm indifferent between $(d,0)$ and (d,r).

The former satisfies:

$$v - 2k > p^r_{a_d}(v-k) + \left(1 - p^r_{a_d}\right)(1-\pi_r)(v-k)$$

which is equivalent to:

$$p^r_{a_d} \geq \overline{p}^r_{a_d} = \frac{v[\pi_d(3-\pi_r) + \pi_r]\pi_r - k(3-\pi_r)(\pi_d + \pi_r + \pi_c\pi_r)}{v[\pi_d(3-\pi_r) + \pi_r]\pi_r - k(2-\pi_r)(\pi_d + \pi_r + \pi_c\pi_r)}$$

The latter satisfies:

$$v - 2k \geq p^r_{a_d}(1-\pi_d)(v-k) + \left(1 - p^r_{a_d}\right)(v-k)$$

which is equivalent to:

$$p^r_{a_d} \geq \underline{p}^r_{a_d} = \frac{k}{(v-k)\pi_d}$$

For all $k \in (0, \underline{k})$, then, any strategy in which the regulator plays a_d with probability

$p^r_{a_d} \in [\underline{p}^r_{a_d}, \overline{p}^r_{a_d}]$ induces the firm to play (d, r) with probability one. These are the infinite payoff equivalent equilibria referred to in point 1 of Proposition 2.

We conclude noting that: $\lim_{k\to 0} \underline{p}^r_{a_d} = 0$, $\lim_{k\to 0} \overline{p}^r_{a_d} = 1$, and $\lim_{k\to \underline{k}} \underline{p}^r_{a_d} = \lim_{k\to \overline{k}} \underline{p}^r_{a_d} =$

$$\begin{aligned} \lim_{k\to 0} \underline{p}^r_{a_d} &= 0 \\ \lim_{k\to 0} \overline{p}^r_{a_d} &= 1 \end{aligned}$$

and:

$$\lim_{k\to \underline{k}} \underline{p}^r_{a_d} = \lim_{k\to \overline{k}} \overline{p}^r_{a_d} = \frac{\pi_r}{\pi_d + \pi_r}$$

Proof of Proposition 3

The proof of Proposition 3 follows immediately from the proofs of Proposition 1 and Proposition 2 if it holds:

$$0 < \underline{k} < k^u < \overline{k} < \frac{v}{2}.$$

The external conditions are satisfied under Assumption2 1 and 2.

It is then sufficient to show that:

$$\underline{k} = \frac{\pi_r \pi_d}{\pi_r + \pi_d + \pi_r \pi_d} < \frac{\pi_r}{1 + \pi_r} = k^u$$

and:

$$k^u = \frac{\pi_r}{1 + \pi_r} < \frac{\pi_r \pi_d}{\pi_r + \pi_d - \pi_r \pi_d} = \overline{k}$$

The former is equivalent to:

$$1 > \frac{\pi_d + \pi_r \pi_d}{\pi_r + \pi_d + \pi_r \pi_d},$$

The latter is equivalent to:

$$\pi_d > 1 - \pi_d,$$

Both of which are satisfied under Assumption 2 as well.

B Additional Tables

Table A.1: Disclosure-Related Terms

GDPR Article	Terms
13/14(1)(a)	`(?i)(verantwortliche|verantwortlich[a-z]+ f[u?]r die datenverarbeitung|kontakt| kontaktdaten)`
13/14(1)(b)	`(?i)(datenschutzbeauftragte|kontakt|kontaktdaten)`
13/14(1)(c)	`(?i)(art[.]?|arti[a-z]+|?)[ ]+6((?:.*){0,8})(ds[-]*g[-]*vo| datenschutzgrundverordnung|gdpr)`
13/14(1)(d)	`(?i)(zweck|rechtsgrundlage|datenverarbeitung|berechtigte[a-z]? interesse)`
13/14(1)(e)	`(?i)(empf[?a]nger)`
13/14(1)(f)	`(?i)(drittland|[u?]bertragung|usa|privacy[ -]shield|datenschutzschild| angemessenheitsbeschluss|schutzniveau)`
13/14(2)(a)	`(?i)(speicherdauer|dauer|speicherung|speicherfrist)`
13/14(2)(b)	`(?i)(recht auf auskunft|auskunftsrecht|auskunft)`
13/14(2)(c)	`(?i)(recht auf widerruf|einwilligung zu widerrufen|widerruf)`
13/14(2)(d)	`(?i)(recht auf beschwerde|beschwerderecht|aufsichtsbeh[?o]rde| landesbeauftragte[a-z]? f[?u]r den datenschutz|recht auf daten[?u]bertragbarkeit|daten[u?]bertragbarkeit)`
13/14(2)(e)	`(?i)(bereitstellung|vertraglich vorgeschrieben|vertrag)`
13/14(2)(f)	`(?i)(automatisierte entscheidungsfindung|profiling)`

Notes: This table summarizes the terms used to identify paragraphs that disclose information pertaining to Art. 13 or 14. The information to be disclosed as listed in Art. 13 is the same as in Art. 14.

Table A.2: Disclosure as the Topic-Weighted Information Volume

		Example 1			Example 2		
	Words	**Topic**	**Factor ϕ_k**	$\phi_k w_{c\|k}$	**Topic**	**Factor ϕ_k**	$\phi_k w_{c\|k}$
Paragraph 1	10	A	2.0	20	A	2.0	20
Paragraph 2	20	B	1.0	20	B	1.0	20
Paragraph 3	30	C	0.5	15	C	0.5	15
Paragraph 4	40	C	0.5	20	A	2.0	80
Total word count	**100**	**Disclosure (Ex. 1)**		**75**	**Disclosure (Ex. 2)**		**135**

Notes: This table illustrates the topic-weighted information volume, using a privacy policy with four paragraphs and their respective word counts. For the overall distinct-topic distribution, we assume $(0.25, 0.25, 0.50)$. For the distinct-topic distribution of disclosing paragraphs, we assume $(0.50, 0.25, 0.25)$. The topic factors are therefore $(\phi_A, \phi_B, \phi_C) = (2, 1, 0.5)$. The two examples differ in the distinct topic for Paragraph 4 ('C' in Example 1, 'A' in Example 2). The unweighted word count of the policy is 100. In Example 1, Paragraph 4 is unlikely a disclosing paragraph: the topic-weighted word count is 75. In Example 2, Paragraph 4 is likely a disclosing paragraph: the topic-weighted word count is 135.

Table A.3: Comparison of Readability with Other Text Corpora

	Obs.	**Word length**	**Sentence length**	**Big words**	**German FRE**	**LIW**
Privacy policy panel	585329	2.16	17.84	0.21	35.98	56.13
		(0.07)	(3.26)	(0.04)	(5.64)	(3.94)
Simple-language news	1594	1.74	10.74	0.04	67.5	39.11
(nachrichtenleicht.de)		(0.12)	(1.8)	(0.03)	(7.28)	(5.42)
Speeches and statements:	1128	1.83	18.16	0.3	54.84	48.05
Angela Merkel		(0.07)	(2.3)	(0.03)	(4.47)	(3.1)
Decisions by German Constitutional Court (BVerfG)	9358	1.96	16.35	0.15	49.27	50.17
		(0.09)	(2.91)	(0.03)	(6.75)	(4.91)
Wikipedia (German)	10000	1.9	20.63	0.12	48.48	53.51
		(0.2)	(14.48)	(0.04)	(18.23)	(15.48)
Wikipedia (English)	10000	1.71	19.78	0.05	60.33	47.8
		(0.16)	(6.57)	(0.03)	(11.58)	(9.31)
GDPR/DS-GVO (Wikipedia)	1	2.1	18.63	0.12	38.35	57.1
GDPR/DS-GVO (official)	1	2.24	40.39	0.18	8.83	81.39

Notes: We report text characteristics (readability factors and the LIW readability score) for our estimation sample and various German-language corpora. *Simple-language news* are all news articles published on nachrichtenleicht.de between November 2019 and June 2023. The *Speeches and statements* by Angela Merkel are from Barbaresi (2019). The *Decisions by the German Constitutional Court* are from Möllers et al. (2021). The Wikipedia pages are from random sample of German pages and their English-language counterparts (accessed in June 2022).

Table A.4: Popularity of Readability Scores

	Google Scholar	
Readability score/index	**Search**	**Citations**
Flesch's Reading Ease Score	~25,000	6069
Gunning's Fog Index	~13,500	2669
Simple Measure of Gobbledygook (SMOG)	~10,600	3143
Lexile Measure	~5300	69
Anderson's Readability Index	4950	242
Automated Readability Index (ARI)	4400	323
Fry Readability	4210	1744
Flesch-Kincaid Readability Score	3990	3698
Simplified Automated Readibility Index	3190	323
Coleman's Readability Formula	2420	134
Coleman-Liau Index	2020	963
The Old Dale-Chall Readability Formula	1050	2473
Fucks' Stilcharakteristik	928	22
The New Dale-Chall Readability Formula	868	1246
Björnsson's Läsbarhetsindex (LIW/LIX)	684	17
Linsear Write	441	1049
Neue Wiener Sachtextformeln (1–4)	366	242
Easy Listening Formula	186	77
Atos Readability	154	2
Wheeler & Smith's Readability Measure	141	44
Farr-Jenkins-Paterson's Simplification of Flesch Reading Ease Score	113	326
EFLAW Readability	109	24
Amstad Verständlichkeitsindex (German FRE)	9	189
Coleman-Liau Estimated Cloze Percent	4	963
Dickes-Steiwer Index	4	42
Danielson-Bryan´s Readability Measure	3	52

Notes: The table reports the number of Google Scholar search results and the number of Google Scholar citations for a variety of readability scores and indices. Numbers are hand collected, accessed March 31, 2023.

C Additional Figures

Figure A.1: Budget for Data Protection Authorities (Total per Capita, in Euros)

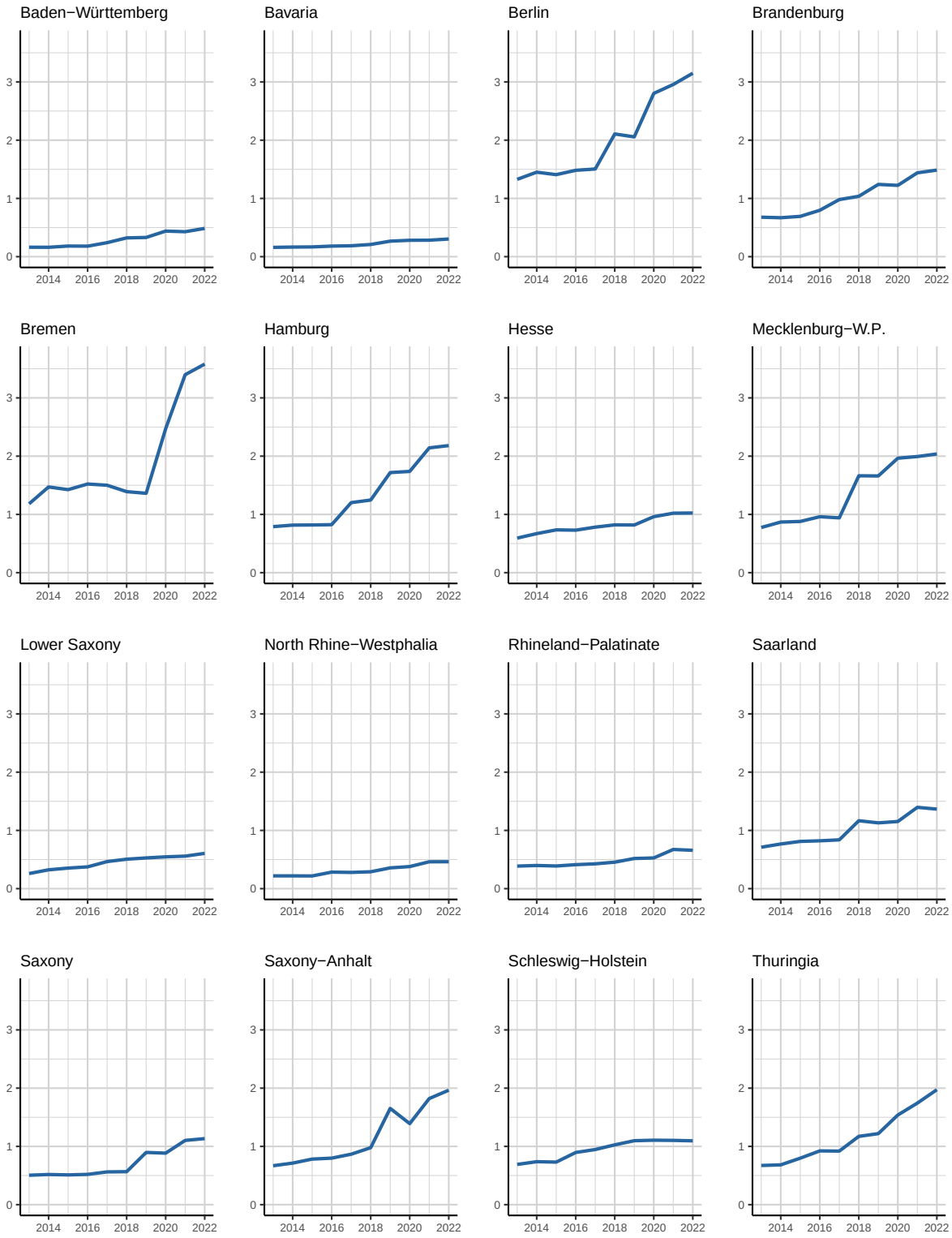


Figure A.2: Budget for Data Protection Authorities (Labor per Capita, in Euros)

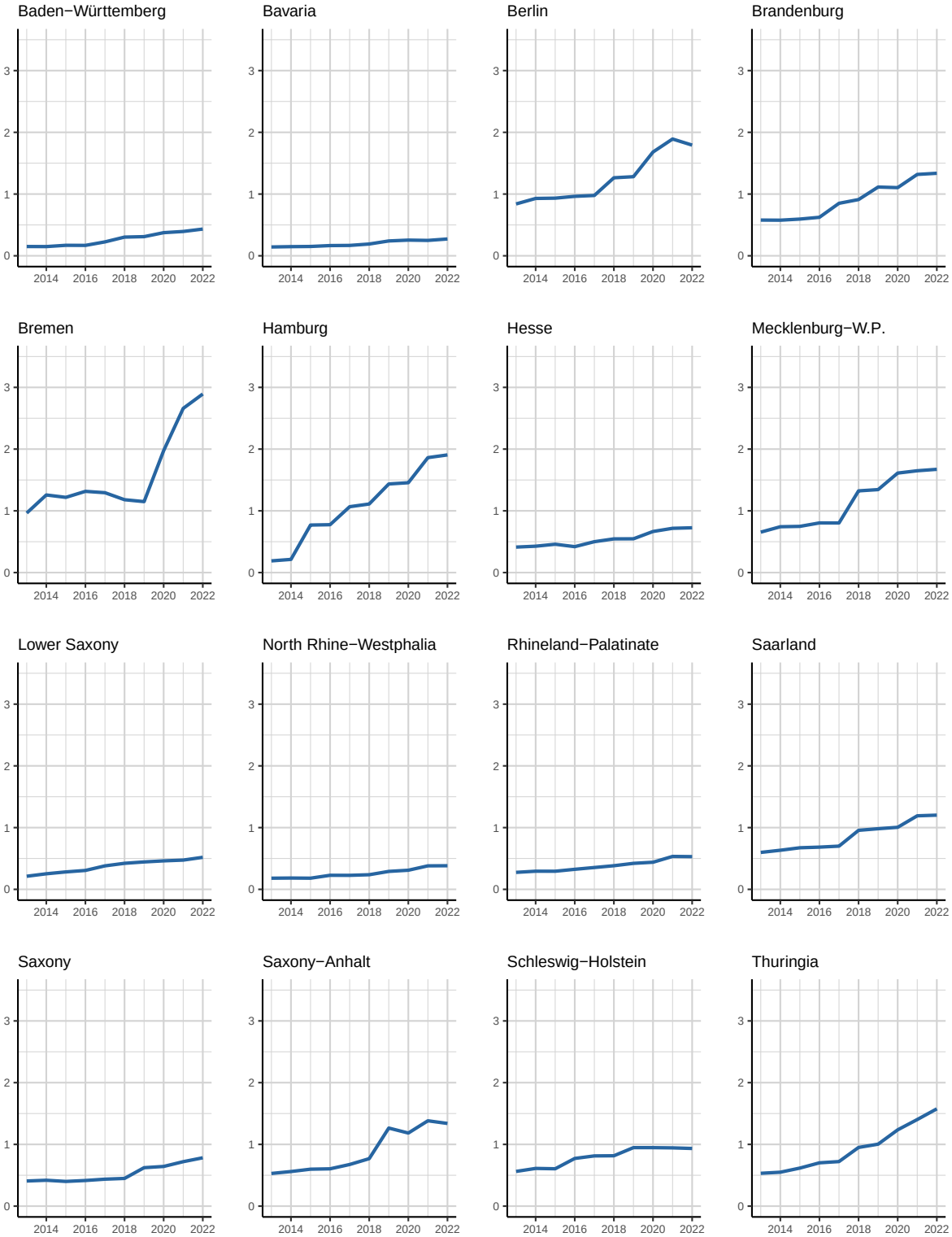


Figure A.3: Budget for Data Protection Authorities (Share of Total State Budget)

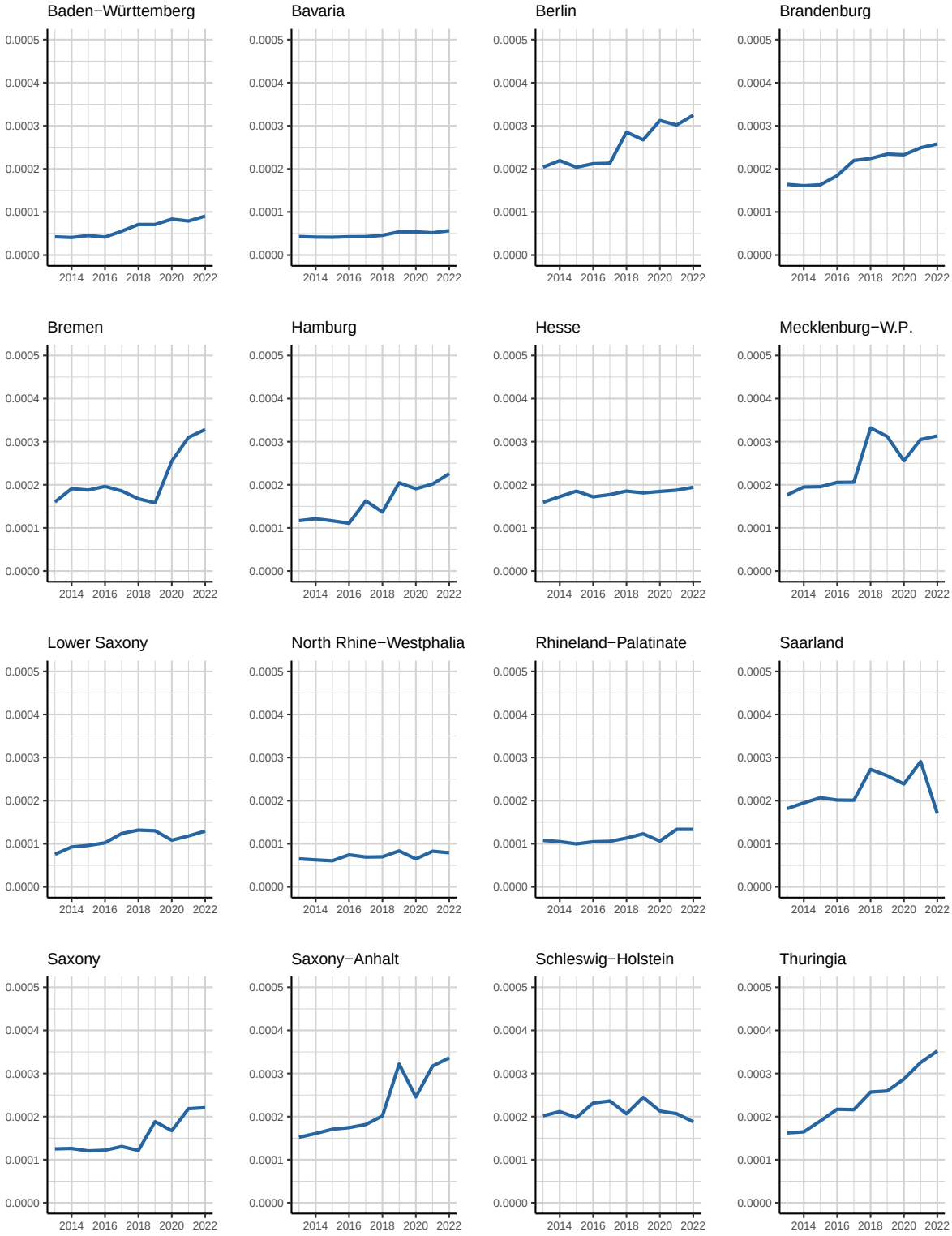
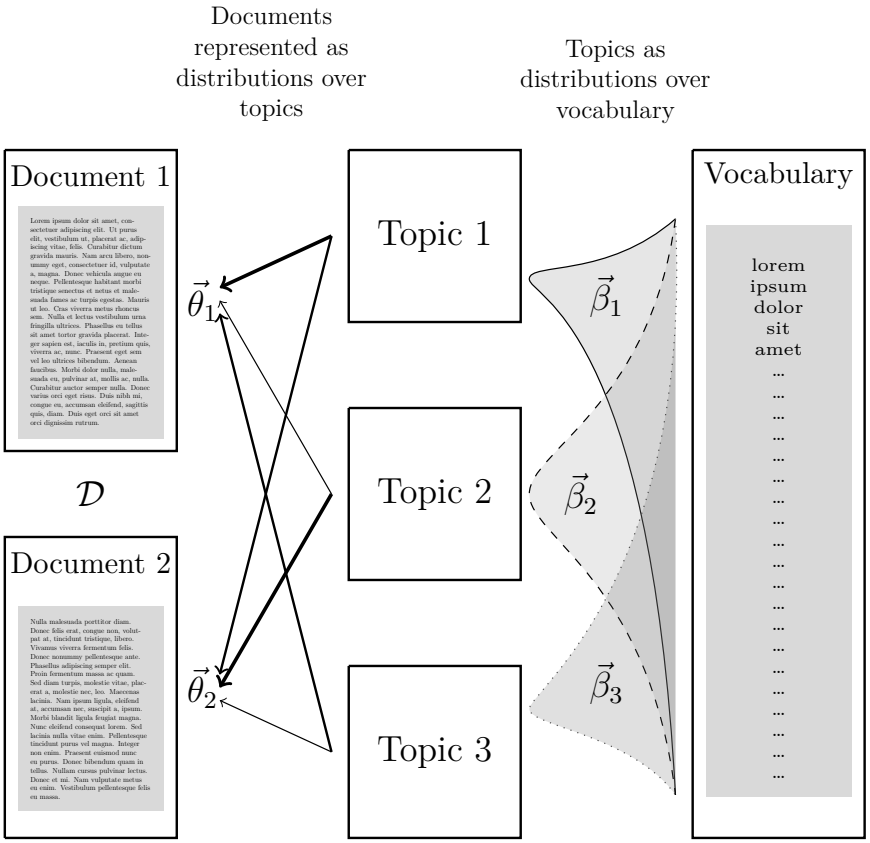


Figure A.4: Probabilistic Topic Models



Source: Ganglmair and Wardlaw (2017)

Table A.5: Data Construction: Most Common URL Patterns

No.	Pattern	Frequency	Share
1	datenschutz/	109642	0.17
2	datenschutz.html	68805	0.28
3	datenschutz	66367	0.38
4	datenschutzerklaerung/	27113	0.43
5	j/privacy	25163	0.47
6	datenschutz.php	12987	0.49
7	datenschutzerklaerung	11554	0.51
8	datenschutzerklaerung.html	9453	0.52

D Construction of the Privacy Policy Panel

The construction of the Privacy Policy Panel begins with an initial sample of 570,000 firm IDs and URLs of the firms' websites containing the privacy policies, taken from the 2019 wave of the Mannheim Web Panel (Kinne and Axenbeck, 2019). We select the privacy policy pages by sampling those URLs from the web panel that contain the term "datenschutz" (the German word for data protection) or "privacy". This section discusses the individual steps of our data construction. All of the relevant code can be found on the official gitlab page https://gitlab.com/MaPPPanel.

D.1 Internet Archive/Wayback Machine Collection

The Wayback Machine is a part of the Internet Archive, an organization founded in 1996 with the intent to preserve the history of the internet by archiving important websites. The organization repeatedly visits websites and stores snapshots of their content for potential future use. A user accessing the Wayback Machine can then search for a specific website and "visit" its historic versions, which can then be scraped and collected as any real time site would. Figure A.11, for instance, shows the screenshot of the homepage of the ZEW Mannheim as it was stored in August 16, 2001 and accessed in July 2023.

The scraping process proceeded as follows: From the 2019 wave of the Mannheim Web Panel, we determine the most common URL patterns used by firms to store their privacy policies. The resulting list (Table A.5) cumulatively takes up 52% of all patterns in the referenced wave. For each firm, we extract the registered URL and look for the archived correspondent page for each quarter between 2014 and Q2 of 2021. A website may have moved its location of its privacy policy over time, e.g., from /datenschutz/ to /datenschutz.html. When the original URL does not return a page, the scraper is instructed to cycle through the most common URLs as per Table A.5 until a match is found. If no match is found again, the observation is left empty; if multiple matches are available, the first one is selected for all ambiguous cases.

The Internet Archive restricts what we are able to capture by what was visited and saved in the Wayback Machine in the first place. It is important that we briefly touch upon the way this process takes place. The Archive uses a series of webcrawlers (both directly controlled by the Archive and by third parties, e.g. Alexa crawls) to visit a large amount of websites and save the content of the pages they visit. The webcrawlers are programmed in such a way that, starting from any page, they follow any links contained in it to enrich the collection.[50] The active collection by webcrawlers happens in programmed waves or "crawls" starting from a list of URLs as initial targets.[51]

The data contained in the Wayback Machine has some inherent bias. Crawlers follow links contained on a visited page: the resulting data might suffer from over-representation of large, public and well connected sites compared to smaller economic agents with lower visibility. The end result of a crawl systematically depends on its starting point: some websites might not appear in different crawls. Older firms might furthermore be over-represented since crawlers often revisit sites already seen in the past. At the same time, more recent crawls appear to be more thorough and widespread than older ones. All of the above shape the composition of our final estimation sample. Overall, we expect our sample to be biased towards larger firms (because they are more likely to be mentioned on other sites) and older firms (because the Wayback Machine might occasionally re-visit already stored sites). Furthermore, we expect a bias towards more consumer-facing industries and especially companies where a website is a part of their core product.

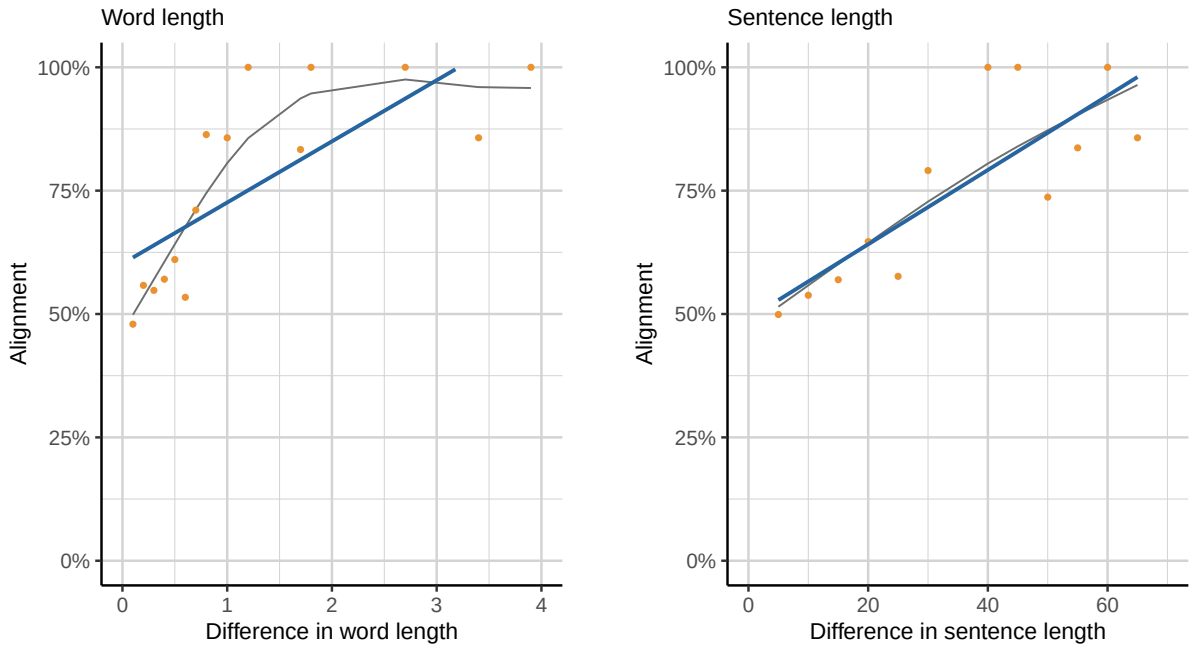
D.2 HTML Parser

We strip from scraped websites the HTML code to preserve the actual policy text. We use a parser (internally developed and tested) that relies on the *readability-lxml* package in Python (https://github.com/buriy/python-readability): an adapted version of the *doc.summary()* function of this package extracts text from the HTML page. We validate our approach (and the performance of the text parser) using a viewer app (internally developed) that displays the stripped text as illustrated in Figure A.12.

[50] This process is usually combined with policies that control and limit the number of links a crawler will follow on a given page, i.e., how "deep" it will go into any single page. This prevents crawlers from ending up in an infinite loop or "getting lost" in one site.

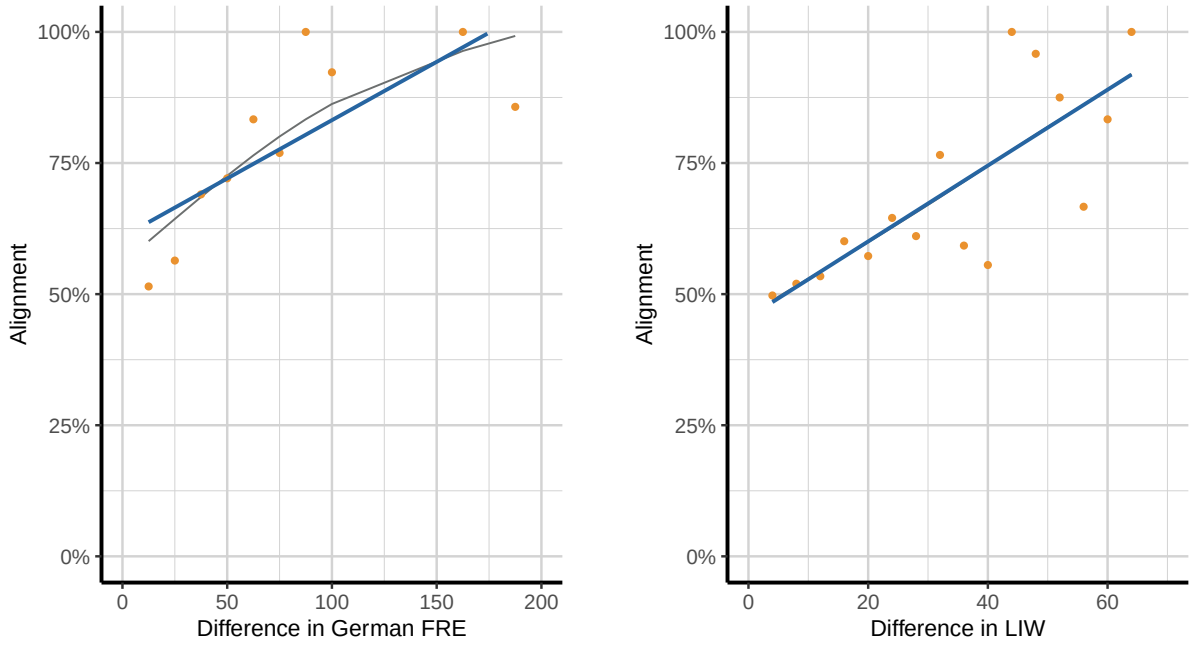
[51] A list of these crawls and respective starting URLs can be found at https://archive.org/details/web.

Figure A.5: Human vs. Factor-Based Assessment of Readability (Factors)



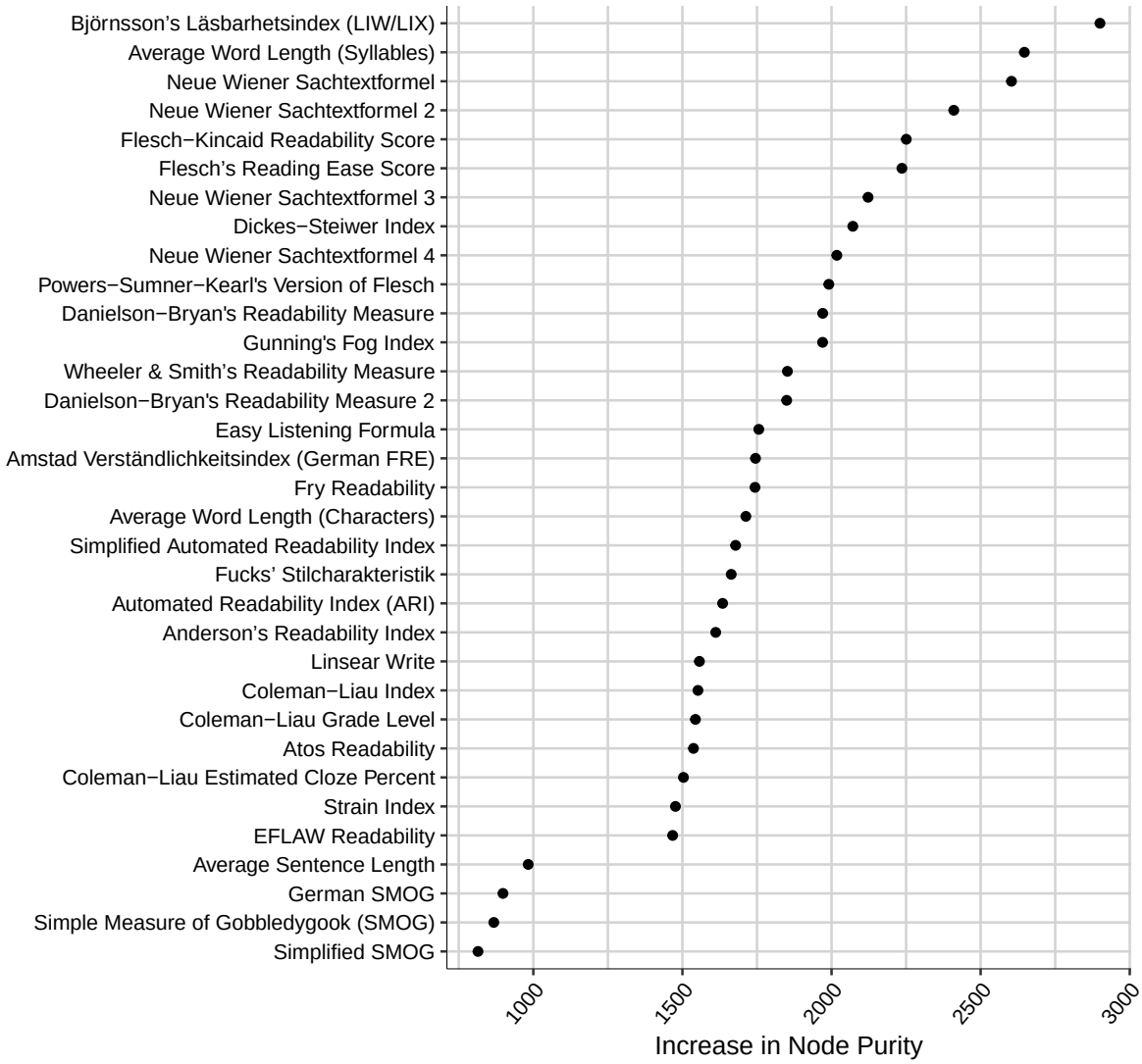
Notes: The figures depict the percentage of text pairs for which the human assessments align with the ranking based on the text pairs' absolute differences in average word length (LHS) and the average sentence length (RHS). Values on the horizontal axis are binned for visual ease. Average alignment for each bin (dots); fitted spline (grey thin line); and linear fit (blue thick line). Dot size does not reflect the number of observations in each bin.

Figure A.6: Human vs. Factor-Based Assessment of Readability (German FRE and LIW)



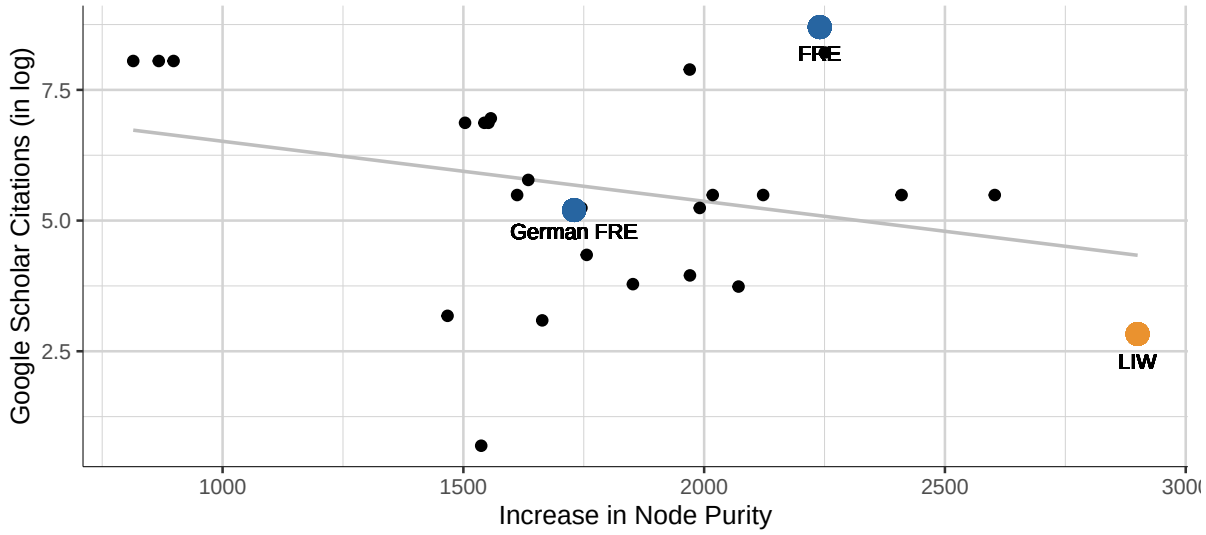
Notes: The figures depict the percentage of text pairs for which the human assessments align with the ranking based on the text pairs' absolute differences in German FRE (LHS) and LIW (RHS). Values on the horizontal axis are binned for visual ease. Average alignment for each bin (dots); fitted spline (grey thin line); and linear fit (blue thick line). Dot size does not reflect the number of observations in each bin. An increase in difference in German FRE of one standard deviation (5.64) increases the alignment by 1.25 percentage points (OLS coefficient: 0.0022, t-statistic: 3.27). An increase in difference in LIW of one standard deviation (3.94) increases the alignment by 2.85 percentage points (OLS coefficient: 0.0072, t-statistic: 4.62).

Figure A.7: Ability to Predict Pair-wise Comparisons



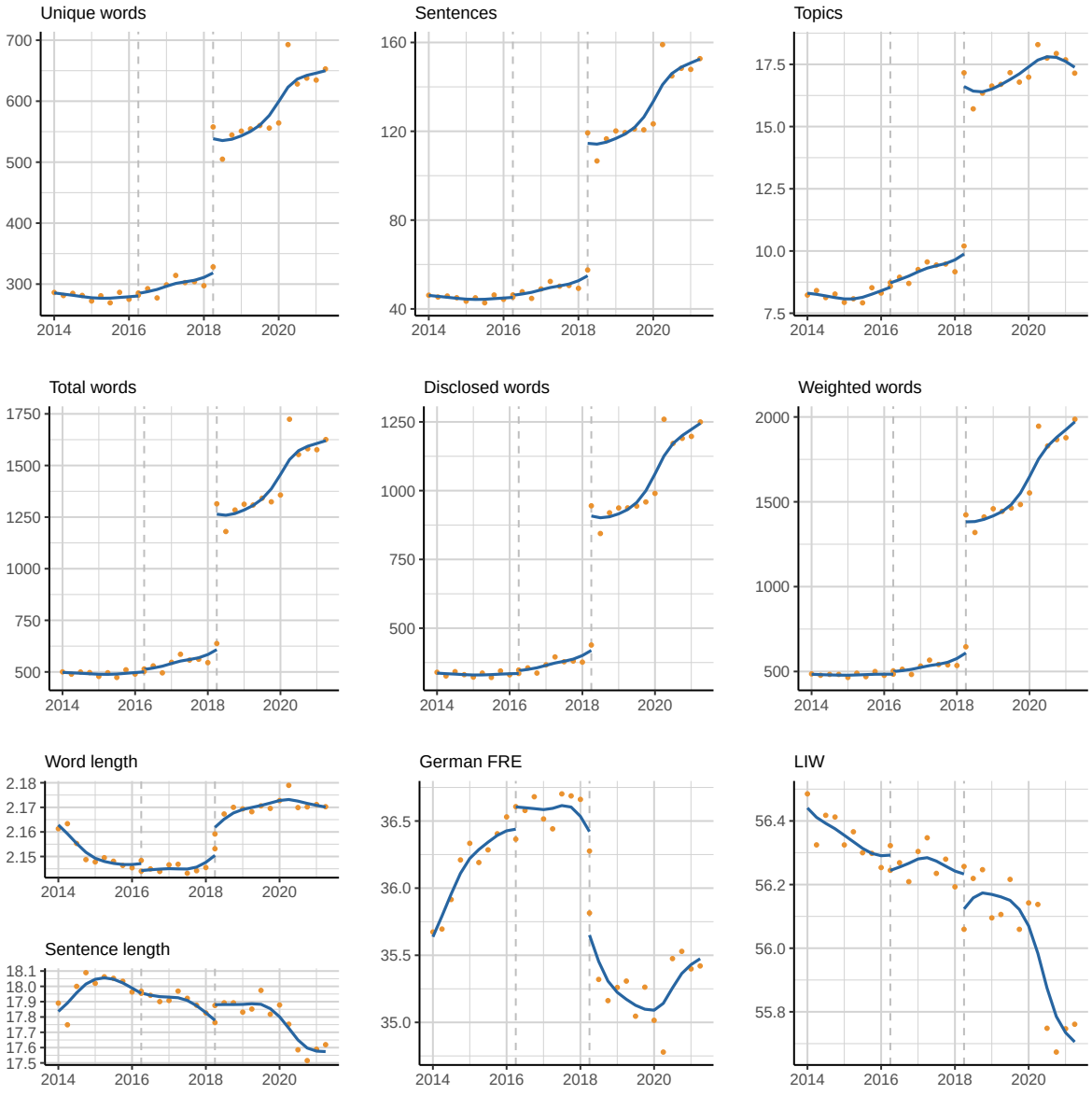
Notes: The figure depicts the performance of readability scores and text-based variables in predicting pair-wise comparisons (following Benoit et al. (2019)). Points to the right with a higher Increase in Node Purity comprise more important variables. Node Purity is a threshold parameter used to determine when node splitting stops and is used to determine the complexity of decision trees. In the present figure, we use 500 trees for the forest.

Figure A.8: Ability to Predict Data vs. Popularity of Readability Scores



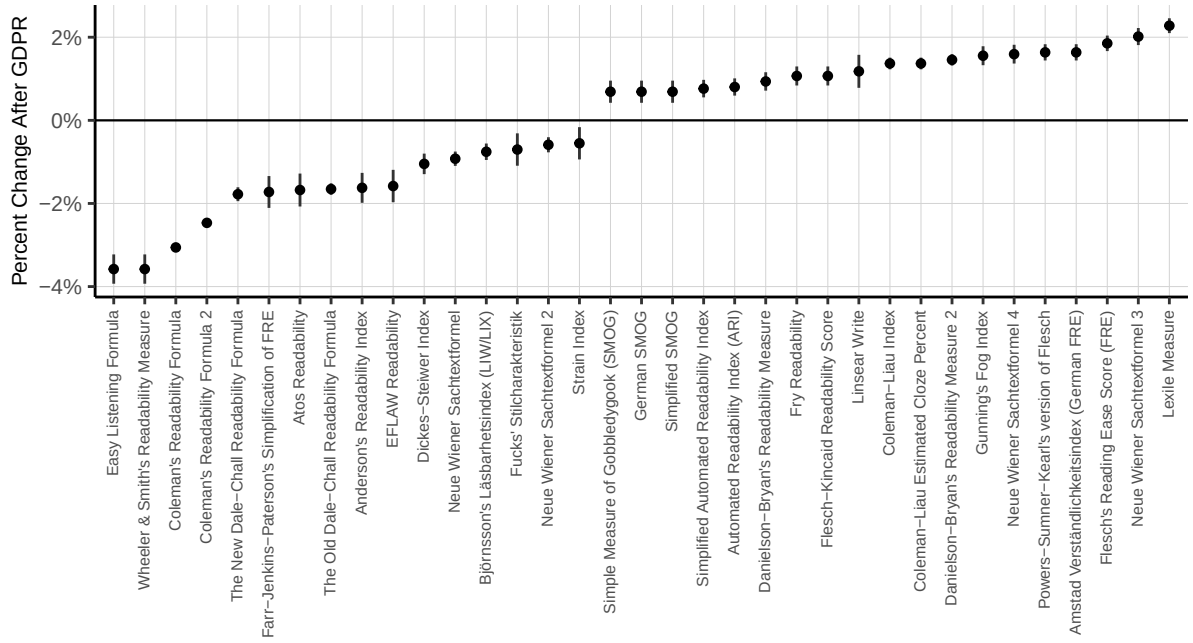
Notes: The figure depicts the performance of readability indices in predicting pair-wise comparisons (following Benoit et al. (2019)) on the horizontal axis and their popularity (Google Scholar Citations) on the vertical axis. We mark our preferred index, the LIW, (in orange), Flesch's Reading Ease Score (FRE) and Amstad's Verständlichkeitsindex (German FRE) (in blue) that has been used in regulation of insurance contract language in the U.S. states of Florida, Massachusetts, Michigan, and Texas.

Figure A.9: Volume, Disclosure, and Readability (Levels)



Notes: This figure presents quarterly averages of policy-level measures for informational volume (panel (a)), disclosure (panel (b)), and readability (panel (c)). Dots represent quarterly averages, the curves are fitted to the data (spline). The vertical dashed lines indicate the GDPR passage in Q2 2016 and GDPR enforcement in Q2 2018. Figure 4 presents the normalized quarterly averages (2014 Q1 = 1).

Figure A.10: Readability Results are Not Robust



Notes: This figure depicts the GDPR-associated change in readability for various readability indices (see Table A.4). We report the GDPR coefficients (and 99% confidence intervals). We have aligned all readability scores so that higher % changes imply less readable privacy policies after the enforcement of the GDPR. The results are without firm-level and industry-level characteristics.

Figure A.11: Data Construction: Screenshot (ZEW Homepage, August 2001)



Notes: This is a screenshot of the homepage of the ZEW Mannheim as viewed and stored on August 16, 2021. This snapshot is from an Alexa Crawl. The URL is https://web.archive.org/web/20010816084954/http://www.zew.de/en/.

Figure A.12: Data Construction: GUI of the Viewer App

