Sumandea-Simionescu, Ioan; Ştefan, Oana; Bercea, Lucian

Working Paper

Soft in name: The uneven normativity of European banking guidelines

SAFE Working Paper, No. 496

Provided in Cooperation with:
Leibniz Institute for Financial Research SAFE

Suggested Citation: Sumandea-Simionescu, Ioan; Ştefan, Oana; Bercea, Lucian (2026) : Soft in name: The uneven normativity of European banking guidelines, SAFE Working Paper, No. 496, Leibniz Institute for Financial Research SAFE, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/343559

Ioan Sumandea-Simionescu | Oana Stefan | Lucian Bercea

Soft in Name: The Uneven Normativity of European Banking Guidelines

SAFE Working Paper No. 496 | August 2026

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

FOUNDATIONS OF LAW AND FINANCE

LawFin Working Paper No. 66

Soft in Name: The Uneven Normativity of European Banking Guidelines

Lucian Bercea | Ioan Sumandea-Simionescu | Oana Stefan

Deutsche Forschungsgemeinschaft - DFG Project FOR 2774
Center for Advanced Studies on the Foundations of Law and Finance | House of Finance | Goethe University
Theodor-W.-Adorno-Platz 3 | 60629 Frankfurt am Main | Germany
www.lawfin.uni-frankfurt.de

Soft in Name:
The Uneven Normativity of European Banking Guidelines

Ioan Sumandea-Simionescu* Oana Ștefan† Lucian Bercea‡

August, 2026

Abstract

European Banking Authority guidelines share a single legal status: formally non-binding, subject to comply-or-explain. This article shows that they share very little else. Applying an expert-coded deontic lexicon to the operative provisions of 85 draft–final guideline pairs (2010–2025), we provide the first measurement of what these instruments are actually made of.

The average final guideline contains roughly twenty obligation-bearing terms per 1,000 words, of which 78 per cent are recommendatory, 18 per cent permissive, and 4 per cent unqualified commands, with recommendatory terms outnumbering commands by roughly twenty to one. But the average conceals a range that runs from instruments containing no command at all — nine of the eighty-five — to one in which commands constitute 43 per cent of all normative language. Instruments carrying identical legal status differ in normative content by an order of magnitude.

We then trace what happens between the consultation draft and the final text. The movement is concentrated with unusual precision. *Shall* falls by 70 per cent in density (0.79 to 0.24 per 1,000 words) and its dispersion across instruments collapses by a factor of 12.9; *must*, which expresses materially similar content, does not move at all (0.72 to 0.67; variance ratio 1.35, $p = 0.58$). A word-level reconstruction of every deontic term in the corpus shows that commands do not descend a register: of the 63 per cent of strong-class occurrences that do not survive, 76 per cent are reworded out of modal form or deleted outright, and only 24 per cent are downgraded to a weaker term. Five instruments enter consultation with commands and leave with none.

None of this is visible to the measurement strategy the literature ordinarily uses. Computed over whole documents rather than operative provisions, our index correlates with itself at $r = 0.19$. We argue that soft law's normativity is constructed at the level of individual words, that it varies far more within a legal category than the category suggests, and that measuring it requires attending to which words, in which parts of which document.

*Leibniz Institute for Financial Research SAFE and Goethe University LawFin. Corresponding author: [sumandea@safe-frankfurt.de]. The author gratefully acknowledge research support from the Center for Advanced Studies Foundations of Law and Finance, funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) – project FOR 2774. All remaining errors are his own.

†The Dickson Poon School of Law, King's College London.

‡Faculty of Law, West University of Timișoara.

Keywords: soft law; European Banking Authority; regulatory text; normativity; deontic modality; comply-or-explain

Contents

1 Introduction

The European Banking Authority issues guidelines that competent authorities must "make every effort to comply with," or else publicly explain why they do not.[1] That single sentence does a great deal of work. It establishes an instrument that binds no one and that no supervisor can disregard silently, and it applies uniformly to every guideline the Authority issues, whatever a particular guideline says.

What guidelines say is the subject of this article. They are written in the grammar of obligation: institutions *shall* maintain a policy, they *should* review it annually, competent authorities *may* require more frequent review. Because the instrument has no formal apparatus for grading its own force — it is non-binding in its entirety, or not at all — these distinctions are carried entirely by syntax. A guideline that wishes to signal that something is required, where elsewhere it merely advises, has one means of doing so, and that means is a word.

This makes soft-law instruments unusually legible to textual measurement, and unusually poorly served by the way that measurement has been done. The literature on EU soft law has established across many settings that formally non-binding instruments bind in practice (Eliantonio et al., 2021; Korkea-aho, 2009; Ştefan, 2013). It has not asked how much, or in respect of what. Answering that requires treating the instrument as internally differentiated, and a single legal status tells one very little about what lies inside.

We report three things.

First, what a guideline is made of. Across 85 final instruments, deontic terms occur at roughly twenty per 1,000 words — one every fifty words — and their composition is heavily weighted toward recommendation: 78 per cent recommendatory, 18 per cent permissive, 4 per cent commanding. This is, to our knowledge, the first quantitative characterisation of the instrument class. The spread is more striking than the average. Nine instruments contain no command whatsoever; one is 43 per cent commands. The interquartile range of command share runs from 1.8 to 8.5 per cent. Instruments with identical legal status, issued by the same Authority under the same article, differ in normative intensity by an order of magnitude.

Second, which words move. Comparing each guideline's consultation draft with its final text, we find movement concentrated in a single term. *Shall* density falls from 0.79 to 0.24 per 1,000 words, a 70 per cent reduction, and the variance of *shall* across instruments falls by a factor of 12.9. *Must* — which in these documents does comparable work — moves from 0.72 to 0.67, with a variance ratio of 1.35 that no test distinguishes from zero change. Removing *shall* from the strong class eliminates the class-level effect entirely. Whatever is happening in the drafting interval, it is happening to one word.

Third, how it happens. Aligning draft and final operative text at word level, we trace the disposition of every deontic occurrence in the corpus. Of 992 draft-side commands, 36.7 per cent survive verbatim — against 73.7 per cent for recommendatory terms. Of those that do not survive, 46.3 per cent are reworded into non-modal form, 29.8 per cent are deleted with their clause, and 23.9 per cent are downgraded to a weaker term. Commands are not moved down a register. They are taken out.

A methodological finding follows from the design and we report it as a result in its own right. Computing the same index over whole documents rather than operative provisions produces values correlating with the operative-text measure at $r = 0.19$ across the corpus. Final EBA guidelines carry a Feedback Statement — absent from drafts by construction, dense with quoted obligation language, and scaling with the number of respondents a consultation attracted. Any comparison of draft and final documents that does not separate operative text from

[1] Regulation (EU) No 1093/2010, Article 16(3).

apparatus is measuring document architecture. This applies well beyond the present corpus.

Section 2 sets out the institutional context and the literature the article speaks to. Section 3 describes the corpus, the lexicon and the measurement decisions. Sections 4 through 6 report what guidelines contain, which terms move, and how. Section 7 reports what does not account for the variation. Section 8 sets out the measurement result. Section 9 draws out the implications.

2 Soft Law, Agency Rulemaking, and the Measurement of Normative Register

2.1 Soft law and the problem of gradated normativity

European soft law has been mapped principally by its functions and its effects. Senden's taxonomy sorts instruments by purpose — preparatory and informative, interpretative and decisional, steering — and remains the standard point of departure (Senden, 2004). The scholarship that followed has been concerned overwhelmingly with a single question: whether instruments that are formally non-binding nonetheless bind in practice.

The answer is that they frequently do. Courts engage with soft instruments in ways that give them practical consequence even while denying them legal force (Ştefan, 2013, 2014). National administrations incorporate them into domestic practice through routes that formal non-bindingness does not predict (Korkea-aho, 2009). Soft and hard instruments are deployed together in hybrid arrangements whose combined effect exceeds what either achieves alone (Trubek et al., 2005). Comparative work across ten Member States has since documented how unevenly that uptake proceeds, and how much it varies by policy field and by the administrative culture receiving the instrument (Eliantonio et al., 2021).

This is a substantial achievement and it has largely settled the binary question. What it has not required, and therefore has not produced, is an account of the internal structure of the instruments themselves.

The omission is not obviously a problem until one notices that soft instruments are not internally uniform. A single guideline may provide in one paragraph that institutions *shall* maintain a policy, in the next that they *should* review it annually, and in a third that competent authorities *may* require more frequent review. Three distinct normative positions have been established within one non-binding text. Nothing in the instrument's formal status distinguishes them; nothing in the taxonomy of soft law registers the difference; and a study asking whether *this instrument* binds in practice has no way to record that the answer differs by paragraph.

The point generalises beyond financial regulation, to any instrument that lacks formal gradation and must construct its own — which is to say, of soft law as a category. Where a regulation can distribute obligation across articles of differing legal effect, and a directive can distinguish result from means, a guideline has one register available and must build every distinction inside it. That is why the linguistics of deontic modality (Palmer, 2001; Tiersma, 1999) bears directly on the legal question: the modal system is the only gradation mechanism these instruments possess.

Two features of the EBA setting make the problem unusually tractable. The gradation is consequential, because comply-or-explain gives the language supervisory bite without giving it legal form. And it is observable, because the Authority publishes each instrument twice.

2.2 Agency rulemaking without rulemaking powers

The EBA's institutional position sharpens the problem considerably.

Established within the post-crisis supervisory architecture, the Authority exercises substantial influence over prudential practice without a general power to make binding rules (Busuioc, 2013; Chiti, 2013). Binding technical standards, which do have legal force, must be adopted by the Commission and are constrained both procedurally and in scope. Guidelines require no such endorsement. They are, in consequence, the instrument through which a great deal of supervisory policy is actually made, less because they are preferred on the merits than because they are available.

The constitutional literature on this arrangement has focused on accountability, on the limits of delegation, and on whether an agency exercising this much influence can be adequately controlled (Moloney, 2011). That debate is important and we do not enter it. Our concern is downstream and complementary: whatever one concludes about the legitimacy of agency soft law, the instruments exist, they are drafted with evident care about their normative register, and that register is revised between publication of the draft and publication of the final text.

Understanding what the revision does is prior to evaluating whether the arrangement is defensible. An assessment of soft-law rulemaking that treats every guideline as an undifferentiated non-binding text is assessing something other than what the Authority produces. If some guidelines command and others merely advise — and Section 4 shows that they do, across a range of an order of magnitude — then the question of whether agency soft law claims too much authority cannot be answered at the level of the instrument class. It has to be asked of instruments, and answering it requires a measure.

2.3 Measuring regulatory language

A parallel literature has developed quantitative measures of regulatory text, largely in the American administrative context and largely for purposes of estimating regulatory burden. The best-known approach counts obligation-bearing terms across the Code of Federal Regulations to construct restrictiveness indices at industry or agency level (Al-Ubaydli & McLaughlin, 2017). Related work in finance has established that general-purpose sentiment dictionaries systematically mis-measure specialised text and that domain-specific lexicons are required (Loughran & McDonald, 2011) — a finding whose logic applies with at least equal force to legal register, where the distinction between *shall* and *should* is categorical.

Three features of that literature motivate the approach taken here.

Obligation counts are typically constructed by the analyst. The set of terms treated as mandatory, and their relative weight, is ordinarily specified by the researcher without external validation. This leaves the strength ordering unexamined at precisely the point where it does the most work, and it makes the measure difficult to challenge or to reproduce. We instead have the lexicon coded independently by three legal experts and report the resulting reliability, including where it is low.

They are computed over whole documents. The implicit assumption is that a regulatory document is a single text of uniform kind. Section 8 shows what that assumption costs when the documents being compared carry different apparatus, and the cost is not small.

They measure levels rather than changes. A restrictiveness index computed on final rules can describe a stock of regulation but cannot say what a drafting process does to it. The comparison exploited here — the same instrument, twice, before and after a bounded interval — is available in any notice-and-comment system and is used relatively little.

Where it has been used, in the American rulemaking literature, the question asked has generally concerned outcomes: whether final rules differ from proposed rules, in whose favour, and whether commenter characteristics predict the difference (J. W. Yackee & Yackee, 2006; S. W. Yackee, 2006). That literature has produced a rich account of influence and a thin one of text. It asks whether a provision changed; it does not ask what changed about it. Our question is the second, and it turns out to have an answer the first would not lead one to expect.

2.4 Guidelines, comply-or-explain, and the drafting interval

The EBA was established in 2011 within the European System of Financial Supervision. Guidelines issued under Article 16 of its founding Regulation are addressed to competent authorities and, in many cases, to institutions directly. Competent authorities must state whether they comply; non-compliance is published. The result is an instrument with the force of a regulation in some hands and of a recommendation in others, and which must therefore establish its gradations internally.

Guidelines are developed by EBA staff and standing committees, published in draft as a Consultation Paper, opened for public comment for a period typically of three months, and issued in final form thereafter. The Authority publishes the responses received together with a Feedback Statement summarising them and giving its reply.

The process has two features that shape the analysis. The published versions differ structurally, to begin with. A Consultation Paper comprises an executive summary, background and rationale, the draft provisions, questions for consultees, and an impact assessment. A final instrument comprises compliance and reporting obligations, background and rationale, the operative provisions, and the Feedback Statement. Comparing whole documents compares objects that are not alike; Section 3 describes how we address this and Section 8 shows what is at stake in doing so.

The interval between the two versions also contains more than consultation. It contains internal review, comment from the Board of Supervisors, legal service revision, and alignment with legislation that may have moved in the meantime. We observe the endpoints of that interval, not the process within it, and throughout, we attribute change to the interval itself; consultation is one of its components and we do not claim to have isolated it. Where we refer to consultation, we mean the interval it bounds.

2.5 Corpus

The corpus comprises 85 draft–final guideline pairs published between 2010 and 2025 — 170 documents. Instruments are keyed on the document identifier printed on the cover page rather than on file or folder naming, which is necessary: two structurally unrelated instruments in this corpus carry the code `GL/2017/16`, one a Joint Committee guideline on wire transfers, the other an EBA guideline on probability of default and loss given default estimation. Six Joint Committee instruments are retained and flagged; their consultation procedure is identical.

Three instruments are excluded. `GL/2020/14` pairs a 2013 consultation paper with a 2020 consolidated text, objects seven years apart that are not versions of one another. `GL/2025/02` pairs a 2025 amending instrument with the 2019 guideline it amends rather than with its own draft. `GL/2020/02`, on loan repayment moratoria, was fast-tracked and has no consultation stage — a fact we return to in Section 7, since it is the Authority's flagship COVID-era instrument. A further eight unpaired or non-document artefacts are excluded.

Seven of the 85 are amending rather than free-standing instruments. For these, operative text is defined as the newly adopted provisions consulted upon in that round, excluding parent-guideline text reproduced unchanged, so that draft and final measure the same object.

3 Measurement

3.1 The lexicon

The deontic lexicon was constructed independently of the analysis. Three legal scholars with expertise in EU financial regulation coded a candidate list of obligation-bearing terms into three strength classes, working separately and without reference to the corpus. The consolidated lexicon contains 32 terms.

The coding procedure was as follows. A candidate list was assembled from three sources: the modal verbs conventionally treated as deontic in the linguistics literature (Palmer, 2001); multi-word obligation formulations recurrent in EU financial legislation (*is required to*, *is expected to*, *is encouraged to*); and negated forms of both. Coding then proceeded in two stages. In the first, each coder assigned every candidate to one of three strength classes independently, working from their own knowledge of EU financial regulatory drafting and without sight of the corpus, its frequency distribution, or any preliminary result; coders did not confer at this stage. In the second, the three met to discuss the terms on which they had diverged and agreed a consensus classification. That consensus was reached unanimously for every term: no term required adjudication by majority, and none was excluded for irreducible disagreement.

The reliability figures below therefore describe the *first* stage. This is the appropriate quantity to report, since it measures how far independent expert judgment converges on the strength of these formulations without discussion, but it is not a property of the lexicon actually used, which carries the full agreement of all three coders.

First-round reliability was moderate: Fleiss' $\kappa = 0.491$, Krippendorff's $\alpha = 0.680$, with 17 of 32 terms assigned identically by all three coders before any discussion. These figures are informative in their own right. Legal experts working independently do not fully agree on how binding a given formulation is, which is precisely the gradation this article sets out to map; a candidate list on which independent coders agreed completely would be measuring something simpler than the phenomenon. That the same coders then reached unanimous agreement in discussion indicates that the initial divergence reflected the difficulty of placing borderline formulations on a scale, and no standing disagreement about what the words mean.

The full lexicon, with each coder's assignment, the consensus class and channel membership, is reproduced in Appendix (to be added). Where disagreement clustered is informative in its own right. The unanimous terms are the unambiguous poles: *shall*, *must*, *cannot* and *is required to* at the strong end, *should* and *is expected to* in the middle, *may* and *might* at the weak end. Disagreement concentrated on two groups. The first is negated middle forms — *should not*, *may not*, *would not* — where coders divided over whether negating a recommendation produces a prohibition (strong) or a negative recommendation (medium). The second is scope and threshold language — *at least*, *at a minimum*, *subject to*, *in accordance with* — where the disagreement is essentially about whether a term that constrains the application of an obligation is itself an obligation. Our three-channel decomposition, described below, is a response to precisely this second problem.

Strong terms create an unqualified requirement (*shall*, *must*, *is required to*, and negated forms). *Medium* terms express expectation without unqualified requirement (*should*, *is expected to*). *Weak* terms confer permission (*may*, *can*, *might*).

The lexicon is decomposed along a second, orthogonal dimension. *Deontic* terms impose or relax obligation directly; *specification* terms allocate responsibility or delimit scope without creating a duty; *conditional* terms qualify the circumstances in which a provision applies. All results below use the deontic channel. One consequence deserves statement: *responsible for* is coded strong but is specification rather than deontic — it identifies who bears a duty rather than imposing one — and occurs 478 times across the corpus. It is excluded from strong-class counts.

We also exclude *will* and *will not*. Both are coded strong, but *will* in these documents is predominantly aspectual rather than deontic: it appears in compliance-notification and reporting apparatus ("notifications will be reported to the EBA by this deadline") describing future states of affairs rather than imposing requirements. At 1,276 corpus occurrences it would otherwise constitute the largest single element of the strong class. All results are reported on the resulting nine-term specification; where the eleven-term figures differ materially we report both.

3.2 Matching and counting

Terms are matched against normalised lowercase text using word-boundary regular expressions, in descending order of term length, with consumption: once a span of text has been matched by a term, it is removed from the string before shorter terms are tested. The effect is that *shall not* is counted once, as a prohibition, and the residual *shall* is no longer available to be matched separately.

This is not a technicality. Consider the sentence *"An institution shall not apply the simplified approach where it may not demonstrate compliance."* Under naive unigram-plus-bigram counting this yields five matches — two strong (*shall not*, *shall*), one medium (*may not*), one weak (*may*) — from two actual deontic operators. The bias is systematic rather than random: it inflates every class in proportion to how often its terms appear in negated form, and in the weak-versus-strong case it assigns a single occurrence simultaneously to the two extremes of the scale. Across this corpus, correcting it moves the draft-to-final change in the composite index by up to 3.80 per 1,000 words for individual instruments.

One further adjustment is required. *May* is both the corpus's most frequent weak term and the name of a month, and EBA guidelines are dense with dates. Date constructions matching *[digit] May*, *May [year]* and *[digit] May [year]* are substituted before matching. Across the 170 documents this affects 97 of 3,256 raw *may* matches; two escape the substitution, both in a single document, and the residual is uncorrelated with year ($r = -0.153$, $p = 0.558$).

3.3 Densities, the composite index, and their limits

Definitions

Let i index instruments and $v \in \{d, f\}$ the draft and final versions. For a set of terms T and version v of instrument i, let $n_i^v(T)$ denote the number of matched occurrences of terms in T under the longest-match-first procedure described above, and w_i^v the number of word tokens in the operative text. The *density* of T is

$$\rho_i^v(T) \;=\; \frac{n_i^v(T)}{w_i^v} \times 1000,$$

occurrences per 1,000 words. The quantity of interest throughout is the change across the drafting interval,

$$\Delta\rho_i(T) \;=\; \rho_i^f(T) - \rho_i^d(T).$$

Three instantiations of T are used. Where $T = \mathscr{D}_S$, the nine deontic strong-class terms, ρ is command density and $\Delta\rho$ the article's principal dependent variable; $\mathscr{D}_M$ and $\mathscr{D}_W$ give the corresponding recommendatory and permissive quantities. Where T is a singleton — $\{shall\}$, $\{must\}$ — the same definitions yield the term-level measures of Section 5. Where T is the union of the three deontic classes, ρ is total deontic density, and the *composition share* of a class is

$$\sigma_i^v(\mathscr{D}_c) \;=\; \frac{n_i^v(\mathscr{D}_c)}{n_i^v(\mathscr{D}_S)+n_i^v(\mathscr{D}_M)+n_i^v(\mathscr{D}_W)}, \qquad c \in \{S,M,W\},$$

which is the quantity reported in Section 4. Note that σ is independent of w_i^v: composition shares do not share a denominator with document length, which is why they are the appropriate measure for describing what an instrument contains.

The composite index

Where a single summary of an instrument's normative force is required we use the weighted composite

$$H_i^v \;=\; \omega_S\,\rho_i^v(\mathscr{D}_S) + \omega_M\,\rho_i^v(\mathscr{D}_M) + \omega_W\,\rho_i^v(\mathscr{D}_W), \qquad (\omega_S,\omega_M,\omega_W) = (2,1,-1),$$

with $\Delta H_i = H_i^f - H_i^d$. The weights encode an ordering and make no cardinal claim: a command counts double a recommendation, and a permission counts against, on the reasoning that permissive language actively relaxes an obligation where silence would merely omit one. Appendix A3 reports ΔH across 96 combinations of weighting and lexicon variant.

We use H sparingly, and it is worth explaining why. Summing across strength classes discards precisely the information the classes were constructed to preserve. Two instruments — one that trades commands for recommendations and one that leaves both untouched — can return identical values of ΔH, since $\omega_S\Delta\rho_S$ and $\omega_M\Delta\rho_M$ enter the sum with the same sign. The composite is informative about aggregate normative force and uninformative about composition, and this article is about composition.

All substantive claims are therefore made at class or term level. This carries an incidental benefit. Class and term densities are raw quantities rather than weighted composites, and for any $\lambda > 0$ the one-sample t statistic computed on $\lambda\Delta\rho$ equals that computed on $\Delta\rho$, since the sample mean and its standard error scale identically and the ratio is unchanged. Weighting cannot affect these results at all, which is a stronger property than surviving a sensitivity check.

Why density

Density rather than raw count is necessary because instruments vary substantially in length — operative sections in this corpus run from under a thousand words to over forty thousand — and $n_i^v(T)$ would principally measure w_i^v. That this is not hypothetical is easily shown: raw medium-class and weak-class counts correlate across instruments at $r = 0.85$, which is the signature of a shared length component rather than of any substantive relationship between the two classes.

Density brings its own constraint, which we observe throughout. Two densities computed over the same document share the denominator w_i^v, and a correlation between $\Delta\rho(T_1)$ and $\Delta\rho(T_2)$ can therefore be induced by variation in document length alone. We restrict correlational claims to quantities that do not share a denominator, and where a between-class relationship is at issue we verify it on raw counts before reporting it. Section 7 records one hypothesis that fails this check and is discarded on that basis.

3.4 Tests of dispersion

Two tests of equality of variance are reported throughout. Levene's test compares the mean absolute deviation from the group mean; the Brown–Forsythe variant substitutes the median, which makes it substantially more robust to the skewed, zero-inflated distributions that density measures on this corpus produce. Where the two disagree we report both and treat the more robust result as governing. The variance ratio — draft variance divided by final variance — is reported alongside as an effect size, since a p-value alone conveys nothing about magnitude.

Because draft and final measurements are taken on the same instruments, the two samples are not independent, and both tests are conservative in this setting rather than anti-conservative: dependence between the samples reduces the effective degrees of freedom available to detect a difference. We note this because it means the significant results below are not inflated by the paired design.

3.5 Delimiting operative text

The index is computed over operative provisions only — the numbered guidelines themselves, excluding executive summaries, background and rationale, questions for consultees, impact assessments, compliance and reporting apparatus, and, in final instruments, the Feedback Statement.

Delimitation proceeded in two stages. A first pass located candidate page ranges by searching each document's table of contents and section headings for the conventional markers of the operative section, and wrote them to a review table with the first and last line of each proposed range. A second pass was manual: each of the 170 proposed ranges was inspected against the document, and the boundary corrected where the automatic proposal had over- or under-reached. Of the 170, 37 required a hand-entered override. Extraction then took a page-slice from the reviewed ranges and performed no heading detection of its own.

Separating the two stages has a practical consequence. Every judgment about what constitutes operative text sits in a table that can be inspected; none is executed at extraction time by rules a reader would have to take on trust. The dependent variable of the study was in this sense built by hand, document by document.

The resulting corpus retains a median 30.5 per cent of final-document pages and 50.4 per cent of draft-document pages. The asymmetry is almost entirely the Feedback Statement, which has no counterpart in a consultation paper. Section 8 reports what including it would have done to the results.

4 Anatomy of a Guideline

4.1 Composition

A final EBA guideline contains 19.8 deontic terms per 1,000 words of operative text — one every fifty words. These are densely normative documents; roughly two per cent of their operative vocabulary consists of terms that impose, recommend or permit.

Their composition is heavily weighted toward the recommendatory register.

Table 1: Composition of deontic language in final EBA guidelines, $n = 85$. Pooled figures aggregate all occurrences corpus-wide; instrument-averaged figures give the mean across instruments.

	Density per 1,000 words	Share (pooled)	Share (instrument mean)
Strong (*shall*, *must*)	0.79	4.0%	6.4%
Medium (*should*)	15.42	78.0%	77.0%
Weak (*may*, *can*)	3.55	18.0%	16.6%
Total	19.75	100%	100%

Recommendatory terms outnumber commands by roughly twenty to one. An EBA guideline is, in the most literal sense, a document of recommendations: commands are a small minority of its normative vocabulary, and permissions outnumber them four to one.

The two columns of shares differ slightly. The pooled figure treats every occurrence in the corpus equally and is therefore weighted toward long instruments; the instrument-averaged figure treats every guideline equally regardless of length. That the strong share is higher on the instrument-averaged measure (6.4 against 4.0 per cent) indicates that shorter guidelines are somewhat more command-dense than longer ones, which is consistent with short instruments being procedural and prescriptive where long ones elaborate methodology. We report both throughout and use the pooled figure when describing the corpus and the instrument-averaged figure when describing a typical guideline.

The instrument class has not previously been characterised in these terms, and the characterisation bears directly on how comply-or-explain should be understood. A competent authority notifying compliance with a guideline is, in the typical case, notifying compliance with a document that recommends far more than it requires.

4.2 Dispersion across instruments

The average is a poor description of any particular instrument.

Table 2: Distribution of strong-class share across final instruments, $n = 85$.

Percentile	0	5	10	25	50	75	90	95	100
Strong share	0.0%	0.0%	0.2%	1.8%	4.2%	8.5%	13.7%	19.4%	42.9%

Nine final guidelines contain no command at all. Their operative provisions are constructed entirely from recommendation and permission. At the other end, `GL/2024/12` devotes 42.9 per cent of its normative vocabulary to commands — an instrument in which requirements outnumber recommendations.

Excluding the tails changes little. The middle half of the corpus spans a fivefold difference in command share, from 1.8 to 8.5 per cent; between the tenth and ninetieth percentiles the ratio approaches seventy.

What these numbers describe, concretely: An instrument at the tenth percentile contains, in a typical five-thousand-word operative section, roughly a hundred deontic terms of which fewer than one is a command; an instrument at the ninetieth percentile, in the same space, contains fourteen. The first states what institutions should do and may do. The second, in addition, tells

them what they must. A supervisor reading the two would have no difficulty telling them apart. The legal classification does not.

All of these instruments carry the same legal status: non-binding, subject to Article 16(3), and generating the same formal obligation on a competent authority to declare compliance. Nothing in the category distinguishes an instrument built entirely from recommendations from one in which commands predominate.

We take this to be the more important of the article's two empirical claims. The category "EBA guideline" is legally uniform and normatively heterogeneous, and the heterogeneity dominates the corpus.

4.3 Drafts and finals compared

Draft instruments are constructed on the same model but are measurably harder. Total deontic density is 20.07 per 1,000 words at draft stage against 19.75 at final; the difference is small but statistically distinguishable from zero (-0.92, $t(84) = -2.28$, $p = 0.025$). The compositional difference is larger than the density difference: strong-class share falls from 6.7 to 4.0 per cent pooled, a relative reduction of two-fifths.

Six drafts contain no command; nine finals do. **Five instruments enter consultation containing commands and emerge containing none.** They are `GL/2015/17`, `GL/2015/19`, `GL/2017/13`, `GL/2023/05` and `UNK/GL/2010/216`.

5 Which Words Move

5.1 *Shall* and *must*

The strong class as a whole declines across the drafting interval: mean change -0.664 per 1,000 words ($t(84) = -2.805$, $p = 0.0063$). Medium and weak classes do not move detectably (medium $+0.006$, $p = 0.990$; weak -0.262, $p = 0.056$).

Disaggregated by term, the movement proves to be almost entirely attributable to one word.

Table 3: Term-level density and dispersion, draft versus final, $n = 85$. Variance ratio is draft variance divided by final variance; values above one indicate narrowing. Levene's test uses the mean as centre, Brown–Forsythe the median.

Term	Mean draft	Mean final	sd draft	sd final	Var. ratio	Levene p
shall	0.792	0.238	1.962	0.547	12.88	<0.0001
shall not	—	—	0.198	0.093	4.49	0.0012
must	0.716	0.673	0.888	0.765	1.35	0.581
cannot	—	—	0.379	0.269	1.99	0.136
is required to	—	—	0.072	0.075	0.91	0.869
must not[*]	—	—	0.033	0.016	4.06	0.185

[*] Ten corpus occurrences across both stages; reported for completeness, too rare to test.

Shall falls by 70 per cent in density and its dispersion across instruments narrows by a factor of 12.9. *Must* — which in EBA drafting expresses materially similar content — falls by six per cent, and its dispersion is statistically indistinguishable between the two stages.

The two terms are comparable in the respects that would otherwise explain the difference. They occupy the same strength class, assigned unanimously by all three coders. They are of comparable frequency at draft stage (0.79 and 0.72 per 1,000 words). Both appear across the full period and across sectors. They are frequently interchangeable in context: a provision requiring institutions to maintain a record can be and is written either way within this corpus. Whatever distinguishes their treatment, it is not class, frequency, period or semantic content.

The remaining strong-class terms behave like *must* rather than like *shall*. *Cannot* narrows by a factor of two, which no test distinguishes from zero; *is required to* does not narrow at all, its variance ratio falling marginally below one. Only *shall not* joins *shall*, which is unsurprising given that it is the same operator negated.

The contrast is not an artefact of the strong class being small. Removing *shall* from the class, while leaving *shall not* and five other terms in place, reduces the class-level variance ratio from 5.50 to 1.396 and both tests to non-significance (Levene $p = 0.362$, Brown–Forsythe $p = 0.408$). The class-level result exists because *shall* is in it.

5.2 Narrowing and decline

A term whose mean falls toward a floor of zero will show reduced variance for that reason alone, and we therefore test the narrowing directly.

Restricting to the 31 instruments with non-zero *shall* at both stages — where no floor is available — the variance ratio is 14.40, higher than in the full sample (Levene $p < 0.0001$, Brown–Forsythe $p = 0.012$). On logged densities, which strip out mean-scaling, *shall* gives a variance ratio of 1.459 ($p = 0.006$) and *must* 1.114 ($p = 0.489$): the contrast survives the transformation.

We are equally clear about what does not hold. Measured as coefficient of variation, dispersion barely moves for either term (*shall* 2.478 to 2.301; *must* 1.240 to 1.137). Relative dispersion is largely preserved. The narrowing is a narrowing in absolute terms, and it accompanies the decline in level. *Shall* is removed in rough proportion to how much of it an instrument contained, with the consequence that final instruments are uniformly low in a respect in which drafts are not.

Direction is consistent with this. Instruments above the draft-stage median move by -1.391; those below move by $+0.065$.

6 How Commands Leave the Text

The density figures establish that commands are removed. They do not establish what becomes of them, and the intuitive answer turns out to be wrong.

6.1 Tracing individual occurrences

Answering the question requires following particular words, which aggregates cannot do, and we therefore align each draft and final pair at token level.

The procedure is as follows. Both operative texts are normalised and tokenised on whitespace. A sequence-matching algorithm then produces an alignment between the two token streams, decomposing the pair into a series of blocks each labelled as identical, replaced, deleted or inserted — the standard output of a longest common subsequence diff, applied to

words rather than lines. Every draft-side deontic occurrence is located by token index and assigned to the block containing it. Its disposition follows from the block's label and, where the block is a replacement, from what the corresponding final-side span contains:

- *Unchanged*: the occurrence falls in a block marked identical. The term survives verbatim in the same textual context.
- *Softened* or *hardened*: the occurrence falls in a replacement block whose final-side span contains a deontic term of, respectively, lower or higher strength class. The provision persists and the operator has moved on the scale.
- *Reworded*: the occurrence falls in a replacement block whose final-side span contains no deontic term at all. The provision persists; the obligation marker does not.
- *Deleted*: the occurrence falls in a block marked deleted. Neither the term nor its clause has a counterpart in the final text.

Two limits of this design should be stated at the outset. The categories are defined by textual correspondence and carry no implication about legal effect. *Reworded* records the absence of a modal operator from the matched final span; whether the requirement it expressed has been relaxed is a separate question. The alignment also operates on the whole operative text and not on individual provisions, so a rewording that substantially restructures its surroundings is recorded identically to one that changes a single word. We return to both points below and provide hand-verified examples in Appendix A1.

The procedure traces 15,427 draft-side deontic occurrences across the corpus, of which 992 are strong-class under the nine-term specification.

6.2 The disposition of commands

Table 4: Disposition of strong-class occurrences. Of 992 draft-side occurrences, 364 (36.7 per cent) survive verbatim; the table gives the disposition of the 628 that do not.

Disposition	Share of non-surviving	Description
Reworded	46.3%	Provision persists; obligation marker does not
Deleted	29.8%	Provision absent from final text
Softened	23.9%	Downgraded to a weaker class

Three-quarters of the commands that do not survive as commands are reworded or removed. Fewer than a quarter descend the modal register — and where they do, the descent is highly directional: 137 of the 150 downgrades (91.3 per cent) land specifically on *should*.

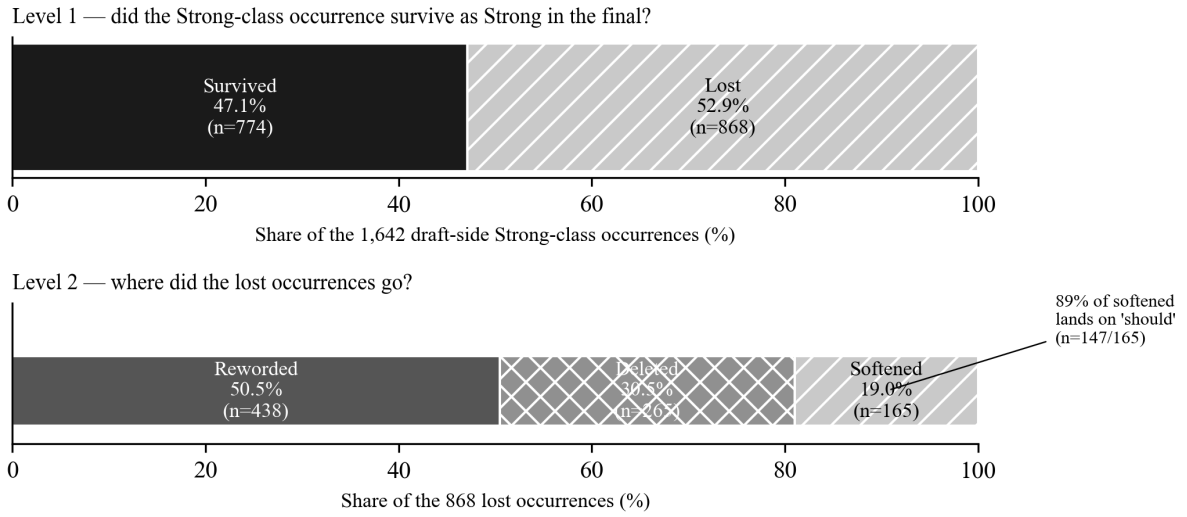

Figure 1: Disposition of strong-class deontic occurrences between draft and final operative text. *Upper panel:* share surviving verbatim as commands. *Lower panel:* disposition of those that do not survive. Figures shown are for the eleven-term specification; the nine-term figures used in the text are 36.7 per cent surviving, and of the non-surviving, 46.3 per cent reworded, 29.8 per cent deleted and 23.9 per cent softened, of which 91.3 per cent land on *should*.

The transition matrix corroborates the class asymmetry. Strong-class terms survive verbatim in 36.7 per cent of cases, against 73.7 per cent for medium terms and 66.0 per cent for weak terms. Commands are twice as likely as recommendations to be absent from the final instrument in the form they took at draft stage.

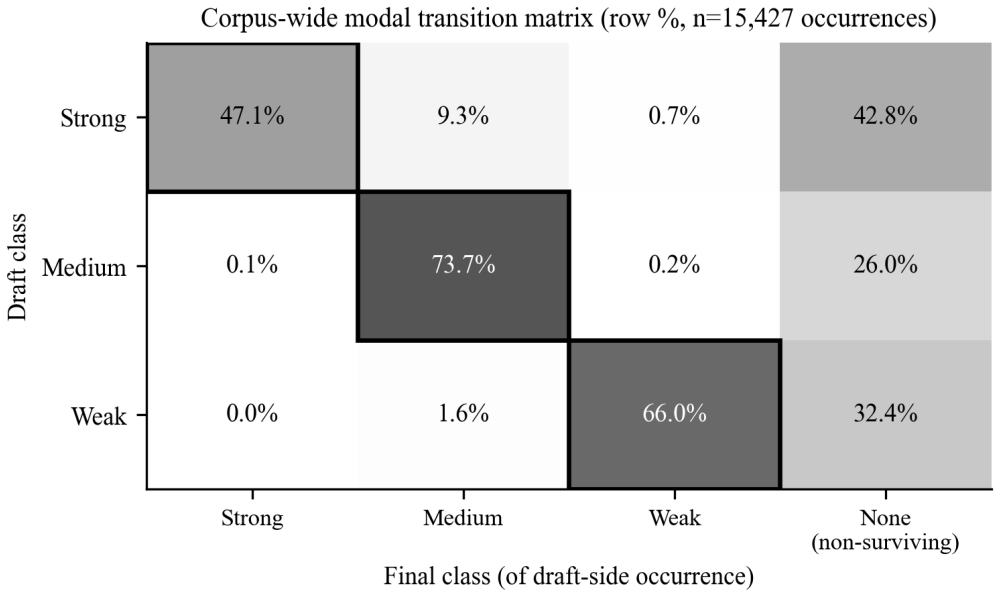

Figure 2: Corpus-wide modal transition matrix, row percentages. The *None* column aggregates occurrences for which no deontic term survives in the matched final span, whether the provision was reworded or deleted. Boxed diagonal cells give verbatim survival. Movement between classes is negligible outside the strong row.

The reworded category is internally varied and we do not overstate what it contains. At one

end are provisions de-modalised close to in place: the margin-of-conservatism requirement in GL/2017/16, framed at draft stage around errors that *cannot* be rectified, reappears in the final text as a medium-class instruction to develop a remediation plan — the same regulatory problem, without the prohibitive framing. At the other end are provisions substantially restructured, where the alignment records a rewording because the surrounding text was rebuilt. Appendix A1 reproduces seven hand-verified pairs spanning both cases, and the boundary between them is a question of legal interpretation that a textual measure locates but cannot resolve.

What the disposition figures establish robustly is the negative claim, and it is the one that matters for how normative gradation in soft law is understood. A literature that treats *shall*-to-*should* substitution as the mechanism by which soft-law instruments are softened is describing the smallest of the three channels. The dominant operations are the disappearance of the obligation marker and the disappearance of the provision.

7 What Does Not Account for the Variation

Normative composition varies enormously across instruments carrying identical legal status. The obvious next question is what organises the variation, and we tested the candidates the literature and the institutional setting suggest. We report them together because the pattern across them is itself the result.

Subject matter does not organise it. Across the eight sectors with at least five instruments, one-way analysis of variance on strong-class share gives $F = 1.118$, $p = 0.365$, with sector accounting for 12.5 per cent of variance. Variation *within* sectors exceeds variation between them: consumer protection, with twelve instruments, spans 0 to 42.9 per cent strong share. Supervisory review and evaluation, with six, spans 1.5 to 3.2 per cent. Two sectors with nothing substantive in common — consumer protection and securitisation — have near-identical compositional profiles, while two supervisory fields sit at opposite ends of the range. Attempts to cluster sectors by compositional profile produce groupings that reorganise under a change of distance metric and are confounded with period ($F = 8.50$, $p = 0.0001$ for year across cluster), and we therefore report no typology.

Time does not organise it. The correlation between publication year and strong-class share is $+0.021$ ($p = 0.847$, $n = 85$). Restricted to instruments consulted upon in 2014 or later, which drops six, draft-side density shows no trend on draft year in any class (strong $p = 0.940$; medium $p = 0.201$; weak $p = 0.069$). Draft-side dispersion is no different before and after 2020 (Levene $p = 0.158$): instruments continue to arrive at consultation as heterogeneous as ever.

Crisis does not organise it. The comparison between the post-Banking- Union and COVID periods gives $t = -0.498$, $p = 0.622$, and does not reach significance in any of 96 combinations of weighting and lexicon variant. It remains null when the corpus is re-dated by consultation year rather than publication year ($t = -1.391$, $p = 0.172$), which reassigns eighteen instruments. The Authority's flagship COVID-era instrument, GL/2020/02 on loan repayment moratoria, was issued without consultation at all — where crisis drove accommodation in this period, it did so by bypassing the process whose interval we observe.

Participation does not organise it. Across twelve specifications — strong-class change and *shall* change as outcomes, raw and logged respondent counts, with and without controls for draft-stage level and final document length — respondent numbers reach significance nowhere ($n = 69$). A falsification test using medium-class change as the outcome is likewise null ($p = 0.926$), so the result is not that heavily consulted instruments are revised more in general.

Length does not organise it. Correlations between change and word count, and between change and operative share, are uniformly weak ($|r| \leq 0.22$).

Nor is the lost obligation relocated elsewhere in the text. Our lexicon distinguishes three channels, and neither of the other two absorbs it. The correlation between deontic and conditional change is -0.130 ($p = 0.235$); the correlation between deontic and specification change is $+0.587$ — positive, which is the opposite of what substitution predicts. A within-channel version of the same hypothesis, in which lost commands are replaced by recommendations, appears in density terms ($r = -0.579$) but does not survive computation on raw counts ($r = -0.090$, $p = 0.414$). The pattern is an artefact of the shared denominator, and we discard it.

Six candidate explanations, none supported. This is a substantive result. Normative composition in this instrument class is not predicted by what the instrument regulates, when it was written, whether it was written during a crisis, how many respondents it attracted, or how long it is. The variation is large and, on everything observable in this corpus, unexplained. Section 9 identifies the most plausible remaining candidate.

8 Measuring the Right Text

We noted that consultation papers and final instruments are structurally dissimilar documents. The consequence is large enough to report as a result.

Computing the index over whole documents rather than operative provisions yields values correlating with the operative-text measure at $r = 0.189$ across the corpus. Excluding five instruments discussed below, the correlation rises to 0.500 — an improvement, and still a moderate correlation between two measures purporting to quantify the same property of the same documents.

The mechanism is the Feedback Statement. It appears only in final instruments; it consists largely of quoted and paraphrased obligation language from respondents and from the Authority's replies; and its length is a function of how many respondents wrote in. Including it means that an instrument attracting many responses is measured differently from one attracting few, for reasons unrelated to its provisions — and, critically, the bias is differential between the two versions being compared, which is precisely the comparison such studies want to make.

Five instruments are affected severely enough to name. In `GL/2024/12`, `GL/2016/06` and `GL/2017/13`, whole-document measurement overstates hardness by between 10 and 26 index points; in `GL/2014/06` and `GL/2021/07` it understates it by 10 and 19 points. In all five the mechanism is the same: a feedback-and-annex share of the final document running from 40.8 to 68.5 per cent, combined with a short operative section whose density is correspondingly unstable.

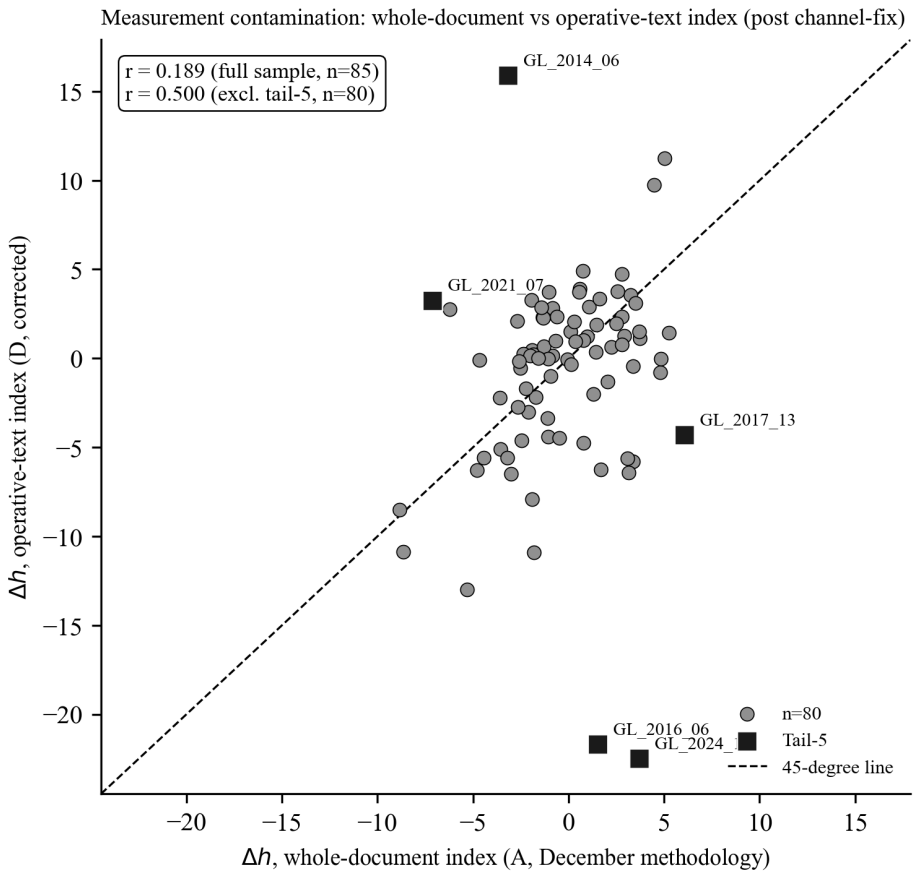

Figure 3: Composite hardness change computed over whole documents (horizontal axis) against the same quantity over operative provisions only (vertical axis), $n = 85$. Perfect agreement would place all points on the dashed 45-degree line. The five labelled instruments carry feedback-and-annex shares of 40.8 to 68.5 per cent of the final document.

The problem is general. Regulatory documents are compound objects, and a text measure computed over them without a scope decision measures document packaging alongside regulatory content. Studies comparing proposed with final rules, in any jurisdiction, face this problem wherever the two versions carry different apparatus.

9 Discussion

9.1 Normativity as composition

The literature on EU soft law has treated softness as a property an instrument has: more or less of it, positioned somewhere between recommendation and regulation. The evidence here suggests something different. Softness is assembled, out of components that behave independently, in proportions that vary enormously across instruments the law treats as identical.

The compositional figures make this concrete. A guideline is, on average, 78 per cent recommendation — but "on average" conceals instruments containing no command at all and one in which commands are the plurality. If normativity were a property of the instrument class, that spread would not exist. It exists because each instrument constructs its own normative position, word by word, and the category imposes no constraint on the result.

This has a consequence for how soft law should be studied. An instrument-level classification — binding or not, hard or soft, Article 16 or otherwise — is carrying almost no information about the object it classifies. The interesting variation is inside the instrument, and reaching it requires measurement at the level of the term.

9.2 The register of command

That the drafting interval acts on *shall* and not on *must* is the article's sharpest empirical result, and we offer an interpretation while being clear that the data do not establish it.

Shall is the term that reads, in EU drafting convention, as a freestanding command issued by the instrument's author. *Must*, *cannot* and *is required to* more often state what a legal framework demands. They describe a requirement's existence where *shall* asserts one. Since the propositional content is frequently identical, the distinction lies in who is understood to be doing the commanding.

Under Article 16(3), a competent authority must state whether it complies. What it is complying with, in practice, is whatever the instrument is legible as requiring. A provision framed with *shall* is legible as a requirement of the guideline. A provision describing what the CRR requires is legible as a statement about the CRR. If the Authority is managing anything here, it is the extent to which its guidelines assert obligations in their own name — a question about the boundaries of its mandate, and a separate one from the substance of its policy.

Two things in the data fit this reading. The substance frequently survives: 46.3 per cent of non-surviving commands are reworded rather than deleted, and the expectation persists in altered form. The effect is also confined to a single term. A general retreat from prescription would move *must* as well, and it does not.

We note the rival account. The plain-language drafting movement has campaigned against *shall* in legal drafting for several decades, and a shift in drafting convention would produce a decline in *shall* with no institutional logic attached. Our data do not support this in the form it would have to take: draft-side density shows no trend on year in any class once the two earliest instruments are set aside, so drafters are not progressively writing softer drafts. A convention account would have to be one in which the convention operates within each drafting interval rather than across the period, which is possible but is a narrower claim than drafting modernisation.

9.3 The unobserved determinant

The most plausible remaining candidate for the variation we document is one this corpus does not contain: the legal basis of each mandate. Guidelines are issued under specific empowerments, and an instrument elaborating a Level 1 provision drafted in the language of obligation plausibly inherits that register, while one elaborating an objectives-based directive does not. This would be orthogonal to sector, to period and to consultation intensity — which is what the corpus requires of any surviving explanation. Testing it requires coding each guideline to its empowering provision and to that provision's own modal register, which is a substantial undertaking and the natural next step.

We are also unable to say what portion of the drafting interval's effect is attributable to consultation specifically. We observe two published texts and the interval between them; that interval contains internal review, supervisory board comment and legal service revision alongside the public consultation. Given that convergence toward a low and uniform level of *shall* is

more readily produced by an institution applying its own drafting discipline than by heterogeneous external respondents commenting on different provisions, we regard internal revision as the more likely locus. But we do not test it.

9.4 Implications

Three follow, in descending order of confidence.

The comply-or-explain regime applies uniformly to instruments that are not uniform. A competent authority declares compliance with a guideline as a whole. Where guidelines range from containing no commands to being 43 per cent commands, that declaration carries very different content in different cases, and the regime provides no means of registering the difference. The point follows from the distribution alone and depends on no causal claim.

Consultation transparency should extend to the text. The Authority publishes a Feedback Statement explaining its responses to comments. It does not publish a comparison of draft and final provisions. Our results indicate that the narrative record and the textual record are separately informative — most of the movement we document is not the subject of any feedback entry — and that a tracked-changes publication would make analysis of this kind unnecessary. The cost would be negligible.

Quantitative measures of regulatory burden computed over whole documents are unreliable. An r of 0.189 between two measurements of the same quantity on the same documents is not a marginal methodological caveat. Any exercise that counts obligation terms across regulatory documents — and several better-regulation and impact-assessment methodologies do — requires a defensible scope decision, and comparisons across document versions require that the versions carry comparable apparatus.

We decline a fourth implication that might be expected. We do not recommend that the Authority use more or fewer commands. Whether de-modalisation weakens a provision substantively is a question of legal interpretation, and our own illustrative pairs show the answer varies case by case. Prescribing modal choice on the basis of a text measure would exceed what the measure supports.

9.5 Limitations

The index has no addressee. "Competent authorities may require institutions to…" is, from the perspective of a regulated institution, a hardening; "institutions may apply a simplified approach" is a relaxation. Our measure treats them identically. Parsing the grammatical subject of each deontic term is the most consequential available extension of this work.

Reworded provisions are classified by grammar, not effect. Whether "institutions ensure X" is substantively weaker than "institutions shall ensure X" is a question our method locates but cannot answer.

One institution, 85 instruments. The EBA's drafting practice may not generalise to other European agencies, and the corpus is small enough that the term-level results rest on modest occurrence counts for all terms but *shall*, *must* and *cannot*.

Sample dependency. `GL/2024/05` amends `GL/2018/08` and `GL/2018/09`, both also in the corpus; these three observations are not fully independent.

10 Conclusion

European Banking Authority guidelines are formally uniform and substantively heterogeneous, and the gap between the two is the article's central finding.

Every instrument in this corpus carries the same legal status: non-binding, subject to Article 16(3), generating the same formal obligation on a competent authority to declare compliance or explain its absence. What they contain varies by an order of magnitude. The typical guideline is built overwhelmingly from recommendation — commands constitute four per cent of its normative vocabulary, and recommendatory terms outnumber them by roughly twenty to one — but nine of the eighty-five contain no command whatsoever, while in one commands are the plurality. Between the tenth and ninetieth percentiles the ratio approaches seventy. Nothing in the legal category records the difference, and nothing observable in the corpus explains it: not subject matter, not period, not crisis, not consultation intensity, not document length.

Between the consultation draft and the final text, that composition shifts in a narrow and specific way. *Shall* falls by seventy per cent in density and its variation across instruments collapses by a factor of 12.9; *must*, which in these documents does comparable work and was coded into the same strength class by all three of our coders, does not move. Remove *shall* from the analysis and the class-level effect disappears entirely. Whatever operates in the drafting interval operates on one word.

Tracing individual occurrences shows what happens to that word: it is removed, not demoted. Of the commands that do not survive as commands, three-quarters are reworded out of modal form or deleted with their provision, and fewer than a quarter descend to a weaker term. Five instruments enter consultation containing commands and emerge containing none. This is a different operation from the one the literature has assumed. A body of work that treats *shall*-to-*should* substitution as the mechanism of softening has been describing the smallest of the three channels through which obligation actually leaves these texts.

We have been deliberate about the limits of the claim. We observe two published documents and the interval between them, and that interval contains internal review, supervisory board comment and legal service revision alongside the public consultation. We do not know which of these produces the pattern, though the character of the pattern — convergence toward a low and uniform level of a single term — is more readily produced by an institution applying its own drafting discipline than by heterogeneous external respondents commenting on different provisions of different instruments. Nor do we know what accounts for the underlying variation in composition; the most plausible remaining candidate, the modal register of each guideline's empowering provision, lies outside this corpus and is where we would look next.

What the article does establish is that these questions can be asked at all, and at what level they must be asked. None of the findings reported here is visible in the instrument's legal status, which is constant across the corpus. None is visible in the composite indices the literature conventionally uses, which sum the strength classes they should be contrasting and which return a null on this data. And none is visible in measurements computed over whole documents, which correlate with the corrected measure at $r = 0.19$ — a figure we report as a substantive result, since it bears on any comparison of proposed and final regulatory text in any jurisdiction where the two versions carry different apparatus.

Soft law's normativity belongs to provisions, not to instruments. It is assembled out of a small vocabulary whose elements behave independently and are revised selectively. What a guideline requires, and how firmly, is therefore a question that has to be asked of its words. Asked that way, it has an answer — and the answer is that instruments sharing a single legal designation differ from one another far more than the designation suggests, and that the

difference is being actively managed in a part of the drafting process no one currently observes.

A1 Illustrative Clause Pairs

The disposition results in Section 6 are produced by an automated word-level alignment of draft and final operative text. Because the substantive reading of a "reworded" disposition depends on what the rewording actually did — a question no automated procedure can settle — we set out below seven hand-verified clause pairs, selected to span the 2010–2024 period and all three non-surviving dispositions, so that readers may assess the classification directly.

Selection excluded every candidate matching the compliance-notification paragraph ("Status of these Guidelines") that appears in near-identical form in almost every instrument. Paragraph renumbering causes the alignment procedure to tag that boilerplate as reworded or softened where no substantive change has occurred; it is excluded here and should be excluded from any further use of the underlying candidate ranking.

Excerpts are reproduced in the lower-cased, normalised form in which the alignment operates. Ellipses mark truncation.

A1.1 Downgraded to a weaker class (*softened*)

A1.1 EBA/GL/2012/02 (stressed VaR), *shall* → *should*. The clearest form of the minority disposition: a single-word substitution, repeated twice in the same passage, with no other change to the sentence.

> **Draft:** …competent authorities **shall** require institutions to comply with the provisions laid down in these guidelines on stressed var. 2. these guidelines **shall** apply to institutions…
>
> **Final:** …competent authorities **should** require institutions to comply with the provisions laid down in these guidelines on stressed var. 2. these guidelines **should** apply to institutions…

A1.2 EBA/GL/2012/03 (CVA risk), *shall* → *should*. The same substitution in the paired instrument issued alongside A1.1, establishing that the pattern is not idiosyncratic to a single drafting team.

> **Draft:** …these guidelines **shall** apply to institutions using an internal model approach (ima) for the purpose of calculating the capital requirements for specific interest…
>
> **Final:** …these guidelines **should** apply to institutions using an internal model approach (ima) for the purpose of calculating the capital requirements for specific interest…

A1.3 EBA/GL/2019/01 (private equity), *shall not* → *should not*. A negated strong term downgraded to a negated medium term, with the surrounding clause otherwise intact. This pair also illustrates why longest-match-first consumption matters: under naive matching this single occurrence would register in both the strong and the negated-form counts.

> **Draft:** …any investments where the institution has the intention to develop a strategic business relationship with the enterprise it has invested in **shall not** be considered as private equity for the purposes of these guidelines…
>
> **Final:** …any investments in which the institution has the intention to develop a strategic business relationship with the enterprise it has invested in **should not** be considered as private equity for the purposes of these guidelines…

A1.2 Obligation marker removed, clause substance retained (*reworded*)

A2.1 EBA/GL/2017/16 (PD and LGD estimation), *cannot* → no equivalent term. The margin-of-conservatism provision is restated without its prohibitive framing. The final text addresses the same subject — deficiencies and uncertainty in risk-parameter estimation — through a positive instruction to develop a remediation plan, in medium-class language, rather than through a statement of what cannot be rectified.

> **Draft:** . . . margin of conservatism aims at addressing all errors that **cannot** be rectified through appropriate adjustment and any other uncertainties related to the estimation of risk parameters. . .
>
> **Final:** . . . following an assessment of the deficiencies or the sources of uncertainty, institutions **should** develop a plan to rectify the data and methodological deficiencies. . .

A2.2 EBA/GL/2014/03 (asset encumbrance disclosure), *shall* → no equivalent term. The draft's reporting instruction for assets encumbered to central banks via emergency liquidity assistance does not survive in its original form. The final instrument restructures the surrounding provision, introducing a qualified exception referring to Article 33 CRR, rather than downgrading the same sentence.

> **Draft:** . . . encumbered to central banks via ela **shall** be reported as unencumbered, such that the sum of encumbered assets corresponds to the institutions' total assets on balance sheet. . .
>
> **Final:** . . . as encumbered, except in certain situations where the institution holds the corresponding covered bonds as referred to in article 33 of the crr. 6. assets placed at facilities. . .

A1.3 Clause absent from final text (*deleted*)

A3.1 CEBS/GL/2010/216, *will* → absent. Consultation-stage procedural text — hearing dates, a scheduled implementation review — with no counterpart in the final instrument. A genuine deletion, but not a deletion of substantive obligation.

> **Draft:** . . . an implementation review **will** be undertaken by the cebs secretariat in the first quarter of 2011. . . a public hearing **will** be held on 9 march 2010 at cebs's premises in london. . .
>
> **Final:** . . . cebs is aware that their implementation **will** encompass not only the amendment of national guidelines, but also supervisory processes. 2. governance mechanisms. . .

A3.2 EBA/GL/2024/05 (STS securitisation amendments), *shall* → absent. The draft's originator-eligibility criteria are framed in strong-class language in consultation-stage explanatory material. No equivalent language survives in the final amending provisions, which are confined to the scope and application of the amendments themselves.

> **Draft:** . . . an originator **shall** be an entity that is authorised or licenced in the union. . . underlying exposures **shall** be originated as part of the core business activity of the originator. . .
>
> **Final:** . . . the amendments to guidelines eba/gl/2018/08 and eba/gl/2018/09 on the sts criteria for abcp and non-abcp securitisation. . . apply to securitisations the securities of which are issued in accordance with terms of agreement adopted after 09/12/2024. . .

A1.4 What these pairs establish, and what they do not

Read together, the seven pairs support the paper's classification unevenly, and we set out the unevenness rather than assert a uniform reading.

The *softened* disposition is unambiguous. A1.1–A1.3 are single-word substitutions in otherwise identical sentences. Where the alignment records a downgrade, a downgrade is what occurred.

The *reworded* disposition is heterogeneous. A2.1 fits the description offered in Section 6 closely: the obligation marker is removed and the substantive expectation is restated in weaker-class language, addressing the same regulatory problem. A2.2 does not fit it as cleanly: the provision is restructured, a qualified exception is introduced, and the draft-to-final correspondence is loose enough that "the clause survives without its obligation marker" understates what changed.

This matters for interpretation. The reworded category, at 50.5% of non-surviving strong occurrences, is the largest single disposition and carries much of the article's argument. If a material share of it consists of substantial restructuring rather than de-modalisation in place, then the claim that consultation removes the grammatical marking of obligation while retaining substance holds for part of that category and not for all of it. We are not able to establish the proportion from the alignment output alone; doing so requires clause-level legal coding of a larger sample, which we identify as the priority extension of this work.

The *deleted* disposition is likewise mixed in kind rather than in reliability. A3.1 and A3.2 are both genuine absences, but neither is the deletion of a live substantive obligation: one is consultation procedure, the other is explanatory framing in an amending instrument. Whether deletions in this corpus disproportionately remove non-operative framing rather than binding requirements is a question the disposition counts cannot answer, and it bears directly on how much normative weight the 30.5% figure should carry.

We report these qualifications because the alternative — presenting only A1.1–A1.3 and A2.1 — would give a misleadingly clean impression of a classification that is, in a substantial minority of cases, a matter of degree.

A2 Supplementary Figures

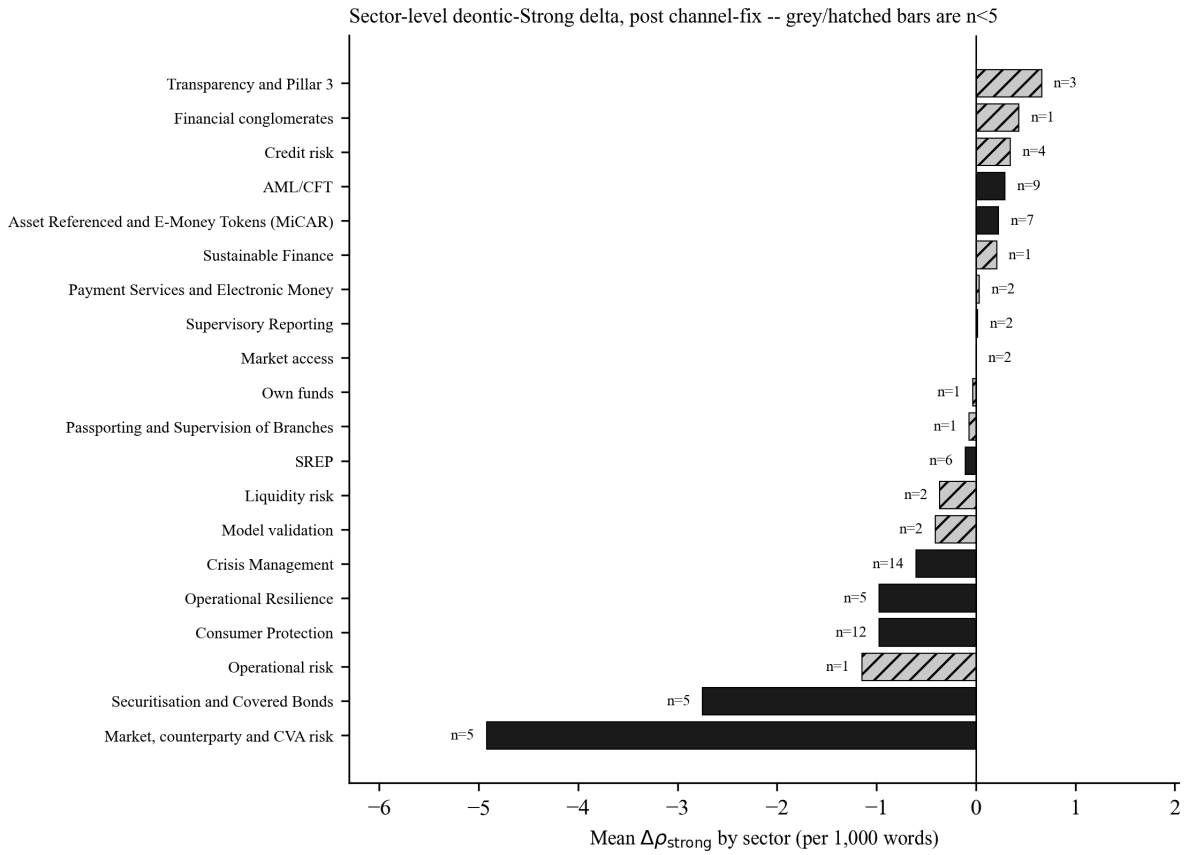

Figure 4: Mean $\Delta\rho_{\text{strong}}$ by primary sector, $n = 85$ instruments across 20 categories, sorted by magnitude with cell sizes annotated. Hatched bars denote sectors with fewer than five instruments. The two largest apparent effects — market, counterparty and CVA risk (-4.92, $n = 5$) and securitisation (-2.75, $n = 5$) — do not survive the checks reported in Section 7: the first is carried by the corpus's only two 2011–2012 instruments and cannot be separated from era, and the second rests on three instruments tied by a single amendment event. The figure is reproduced here so that readers may see the ranking the data produce and the cell sizes on which it rests; it is not offered in support of a sector effect.

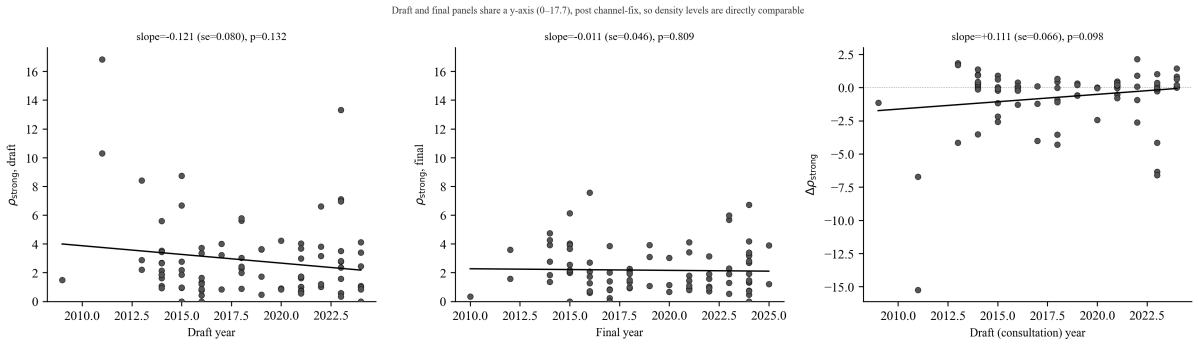

Figure 5: Strong-modal density against year, $n = 85$. *Left:* draft-side density by consultation year (slope -0.121, s.e. 0.080, $p = 0.132$). *Centre:* final-side density by publication year (slope -0.011, s.e. 0.046, $p = 0.809$); the left and centre panels share a y-axis so that levels are directly comparable. *Right:* the consultation effect itself by consultation year (slope $+0.111$, s.e. 0.066, $p = 0.098$). No slope is significant. The pattern — drafts drifting down, finals flat, the consultation effect shrinking — is directionally consistent with the drafting-convention account discussed in Section 9, and this corpus cannot separate that account from a consultation effect.

A3 Weighting and Lexicon Sensitivity

The composite index in Section 3 fixes weights at $2/1/-1$ and uses the consolidated expert lexicon. Both choices are open to challenge, and this appendix reports what happens when they are varied. The class-level results on which our substantive claims rest are, as noted in Section 3, invariant to weighting by construction, so what follows bears on the composite and on the period comparison rather than on the strong-modal finding.

A3.1 The grid

We re-estimate the post-Banking-Union-versus-COVID comparison across a grid of six strong weights (1, 1.5, 2, 2.5, 3, 4) crossed with four weak weights (0, 0.5, 1, 2), with the medium weight fixed at 1, and across four lexicon variants: the consolidated expert lexicon as described in the paper; the same lexicon as implemented in code, which differs in the treatment of a single term; and two broader variants incorporating additional obligation-adjacent terms proposed during expert coding but not retained in the consensus.

The comparison fails to reach conventional significance in *all 96 combinations*. The smallest p-value anywhere in the grid is 0.209, at the most extreme weighting tested (strong weighted four times medium, weak weighted zero) and on the broadest lexicon; at the specification reported in the paper it is 0.894. There is no weighting and no lexicon on which a COVID-period effect appears.

Table 5: Post-Banking-Union versus COVID comparison across the weighting grid, by lexicon variant. Each cell summarises 24 weight combinations.

Lexicon variant	Significant at $p < .05$	p range	Mean ΔH range
Consolidated (as reported)	0 of 24	0.687–0.983	-2.25 to $+0.14$
Consolidated (as coded)	0 of 24	0.687–0.926	-2.25 to -0.10
Broad consensus	0 of 24	0.209–0.907	-2.50 to -0.07
Broad verbatim	0 of 24	0.316–0.976	-2.34 to $+0.13$

One feature of Table 5 deserves note. The mean composite change is negative in 94 of the 96 cells, and the two exceptions sit at the extreme of the weak-weight range. The direction of the aggregate is therefore stable even though its magnitude is not distinguishable from zero: the corpus does not harden under any specification we tested, it simply fails to soften detectably.

A3.2 Do the lexicon variants measure the same thing?

A grid is only reassuring if the variants are genuinely alternative measures of one construct rather than measures of different constructs. We therefore report their agreement directly. Instrument-level composite deltas correlate at $r = 0.996$ between the two consolidated variants, at 0.937–0.938 between the consolidated and broad-consensus variants, and at 0.925 between the consolidated and broad-verbatim variants; the two broad variants correlate at 0.972.

Correlation is a permissive standard, so we also report sign agreement, which is stricter and more informative for a difference measure. Relative to the consolidated lexicon, the direction of the estimated consultation effect flips for 2.4% of instruments under the as-coded variant, 10.6% under broad consensus, and 15.3% under broad verbatim.

The last figure is the one we would flag to a reader. A lexicon that admits obligation-adjacent terms rejected at the consensus stage changes the sign of the estimated effect for roughly one instrument in seven. That is not a threat to the aggregate results, which are null under every variant, but it is a caution against instrument-level interpretation of composite deltas, and it is a further reason to prefer the class-level measures used in the body of the paper — which rest on terms whose classification was substantially agreed rather than on the outer boundary of the lexicon, where agreement is weakest.

References

Al-Ubaydli, O., & McLaughlin, P. A. (2017). Regdata: A numerical database on industry-specific regulations for all united states industries and federal regulations, 1997–2012 [The best-known restrictiveness-by-word-count approach; a useful foil, since it counts obligation terms without a scope decision.]. *Regulation & Governance*, *11*(1).

Busuioc, M. (2013). Rule-making by the european financial supervisory authorities: Walking a tight rope [Directly on point for the EBA's soft-law powers and their constitutional limits.]. *European Law Journal*, *19*(1).

Chiti, E. (2013). European agencies' rulemaking: Powers, procedures and assessment. *European Law Journal*, *19*(1).

Eliantonio, M., Korkea-aho, E., & Ştefan, O. (Eds.). (2021). *Eu soft law in the member states: Theoretical findings and empirical evidence* [SoLaR network; empirical work across ten Member States covering competition, financial regulation, environment and social policy. The financial-regulation chapters are the closest existing empirical companion to this article.]. Hart Publishing.

Korkea-aho, E. (2009). Eu soft law in domestic legal systems: Flexibility and diversity guaranteed? *Maastricht Journal of European and Comparative Law*, *16*(3).

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks [Cite for the general point that general-purpose lexicons mismeasure domain text – which is the argument for expert coding in Section III.A.]. *Journal of Finance*, *66*(1).

Moloney, N. (2011). The european securities and markets authority and institutional design for the EU financial market [VERIFY volume/issue. Companion pieces cover the EBA.]. *European Business Organization Law Review*, *12*.

Palmer, F. R. (2001). *Mood and modality* (2nd ed.) [Standard reference for the deontic/epistemic distinction on which the strength classes rest.]. Cambridge University Press.

Senden, L. (2004). *Soft law in european community law* [The foundational taxonomy: preparatory/informative, interpretative/decisional, steering. Cite for the definitional apparatus in Section II.]. Hart Publishing.

Ştefan, O. (2013). *Soft law in court: Competition law, state aid and the court of justice of the european union* [Textual and doctrinal analysis of 696 documents; the standard reference for judicial engagement with EU soft law.]. Kluwer Law International.

Ştefan, O. (2014). Helping loose ends meet? the judicial acknowledgement of soft law as a tool of multi-level governance. *Maastricht Journal of European and Comparative Law*, *21*(2).

Tiersma, P. M. (1999). *Legal language*. University of Chicago Press.

Trubek, D. M., Cottrell, P., & Nance, M. (2005). "soft law," "hard law," and european integration: Toward a theory of hybridity [VERIFY: also published in revised form in a Hart edited volume.]. *Jean Monnet Working Paper*, (02/05).

Yackee, J. W., & Yackee, S. W. (2006). A bias towards business? assessing interest group influence on the u.s. bureaucracy. *The Journal of Politics*. https://doi.org/10.1111/J.1468-2508.2006.00375.X

Yackee, S. W. (2006). Sweet-talking the fourth branch: The influence of interest group comments on federal agency rulemaking. https://doi.org/10.1093/JOPART/MUI042

LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe

Recent Issues

No. 495	Renée Adams	The Value(s) of Women Artists
No. 494	Dvir Aviam Ezra	The Legal Gray Zone of Proxy Voting Policies
No. 493	Ioan Sumandea-Simionescu	Paying for Prudence? Empirical Evidence on Remuneration in Supervisory Banking Functions During Crisis
No. 492	Andrea Poinelli, Loriana Pelizzon, Davide Tomio, Benoît Nguyen, Tobias Linzert	Elastic in Cash, Inelastic in Repo: Hedge Funds in the Treasury and Repo Markets
No. 491	Konstantin Egorov, Vasily Korovkin, Alexey Makarin, Dzhamilya Nigmatulina	Risk of Sanctions
No. 490	Claes Bäckman	One Advisor for the Whole World? Cross-Country Evidence on Financial Advice from Large Language Models
No. 489	Aras Canipek, Axel Kind, Lubomir Litov, Jiri Tresl	Bankruptcy Law and Firm Size
No. 488	Aras Canipek	Bankruptcy Law Penalties and Board Independence
No. 487	Salvatore Federico, Andrea Modena, Luca Regis	Coordinating Bank Dividend and Capital Regulation
No. 486	Ulrich Bindseil, Svetla Daskalova, Richard Senner	Gold and the External Wealth of Nations
No. 485	Rainer Niemann, Anna Rohlfing-Bastian	Carbon Taxes and ESG Compensation
No. 484	Vanya Horneff, Raimond Maurer, Olivia S. Mitchell	Defaulting 401(k) Assets into Payout Annuities For “Pretty Good” Lifetime Incomes
No. 483	Luca Enriques, Casimiro A. Nigro, Tobias Tröger	Templates in the EU Inc. Regulation Proposal
No. 482	Alexander Morell, Daniel Schellenberg	Regulating Third-Party Litigation Funding? Weighing the Arguments in the Light of Extant EU Consumer Protection Safeguards