



International Journal of Corporate Governance and Public Policy

Journal homepage: <https://trpub.org/journal.php?id=HAZ4qZ4qPA>

TRPub

IJCGPP

International Journal of
Corporate Governance
and Public Policy

Vol. 1 Issue 1 2026

Methodology Article

Open Access

Constructing a Text-Based Index of Lobbying Transparency From Public Consultation Records

Sigrid Harkness^{1*}, Kevin Jovanovic², Minjun Nwosu³

¹Department of Business, Patagonia Institute for Advanced Business, United Kingdom

²Department of Business, Niger Academy of Business, Uganda

³Department of Business, Rhone Institute for Advanced Business, Trinidad and Tobago

Article Info

Volume 1, Issue 1, 2026

[doi: 10.62715/ijcgpp.2026.v1.i1.pp27-32](https://doi.org/10.62715/ijcgpp.2026.v1.i1.pp27-32)

Received: 03 February 2026

Accepted: 11 April 2026

Published: 20 June 2026

Abstract

Lobbying registers record who lobbies and, in some countries, how much is spent, but they exist in a minority of jurisdictions and say little about what is disclosed at the moment an organisation tries to shape a rule. Public consultations offer a second source. When a regulator invites comment on a proposal and publishes the responses, each submission is a lobbying act with a text attached, and the text shows how far the submitter reveals who it is, whom it speaks for, who funds it and what evidence it relies on. This article proposes a procedure for turning such records into an index of lobbying transparency. The procedure has seven steps: defining the construct and its dimensions, assembling the corpus, cleaning the texts and linking submitters, hand-coding a sample against a written codebook, scaling the coding with dictionaries and a supervised classifier, aggregating indicator scores into an index, and validating the result before release. The article explains the choices at each step and the records to keep. The procedure is a proposal built from the literature on text analysis, measurement validity and composite indicators. It has not been applied or validated here, and no index values are reported.

Keywords: Lobbying transparency, Public consultation, Text as data, Composite index, Measurement validity, Interest groups

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

1. Introduction

International guidance treats transparency as the first condition of legitimate lobbying. The OECD recommendation of 2010 asks countries to provide enough information about lobbying for the public to scrutinise it (OECD, 2010), and the organisation's later stocktake found that many members still lacked the means to do so (OECD, 2021). Comparative research has responded by scoring the stringency of lobbying laws country by country (Chari et al., 2010). Indices of that kind measure rules. They do not measure what lobbyists in fact disclose, and they are silent where no lobbying law exists.

Consultation records fill part of that gap. Notice-and-comment rulemaking in the United States, the European Commission's public consultations and similar procedures elsewhere generate large, public collections of submissions from firms, trade associations, professional bodies, civil society groups and

* Corresponding author: Sigrid Harkness, Department of Business, Patagonia Institute for Advanced Business, United Kingdom. E-mail: sigrid.harkness@patagonia.edu

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

individuals. Political scientists have used these texts to study influence, by comparing what submitters asked for with what the final rule contained (Yackee and Yackee, 2006; Klüver, 2013). They have been used far less to study disclosure. Yet a submission either says whom the author represents or it does not, and work on corporate funding of non-profit commenters shows that undeclared ties are of real consequence (Bertrand et al., 2021).

This article proposes a procedure for constructing a text-based index of lobbying transparency from consultation submissions. It is written for researchers and for integrity bodies that hold such records. The procedure is described and justified step by step; it is not applied, and the article reports no results.

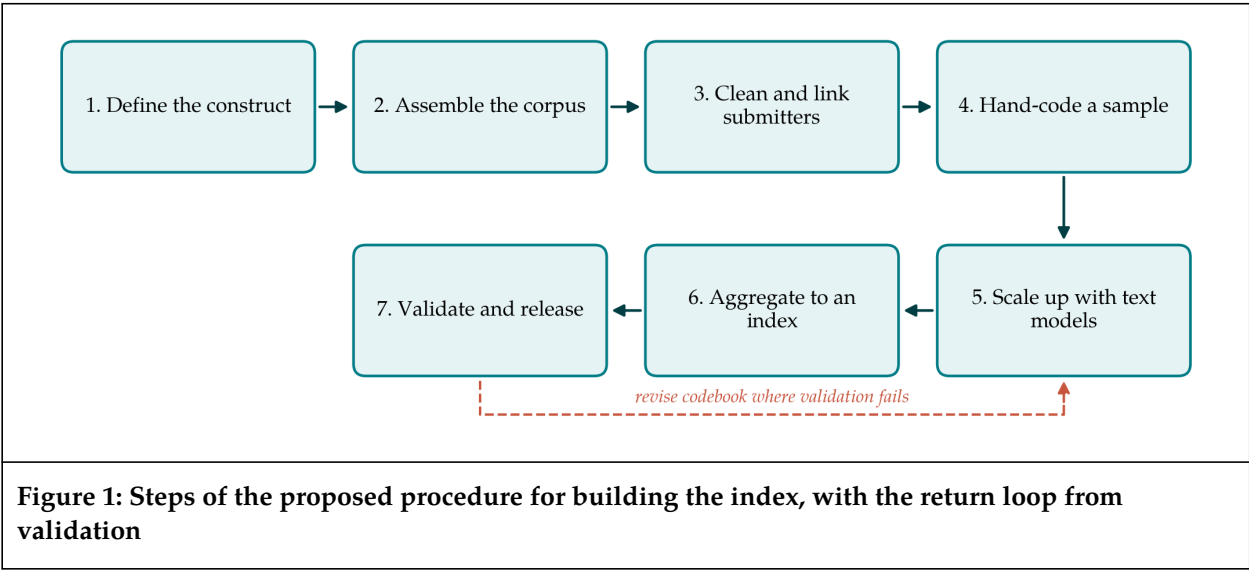
2. Background

Text analysis in the social sciences rests on a small number of principles that shape the design. Grimmer and Stewart (2013) state the most important: automated methods are models of language that are always wrong in some respect, they augment careful reading and do not replace it, and every application must be validated for its own purpose. The methods fall into two families. Dictionary methods count words or phrases chosen in advance; they are transparent, and they fail when terms are used in a domain other than the one the dictionary was built for, as Loughran and McDonald (2011) showed for financial text. Supervised methods learn from documents that people have coded and extend the coding to the rest (Gentzkow et al., 2019).

Two precedents show that an index can be built from text. Baker et al. (2016) constructed an index of economic policy uncertainty from newspaper articles and validated it with a large human audit. In interest group research, Klüver (2009) tested quantitative text methods on consultation submissions against hand coding. The literature on composite indicators adds a separate set of requirements: a stated theoretical framework, defensible normalisation and weighting, and analysis of how sensitive the result is to those choices (OECD and JRC, 2008).

3. The Proposed Procedure

The procedure has seven steps, shown in Figure 1. They are described in order, though in practice the fourth and seventh send the researcher back to the first.



3.1. Step 1: Define the construct

Measurement begins with a concept and proceeds through a systematised definition to indicators and scores (Adcock and Collier, 2001). Lobbying transparency is here defined as the extent to which a submission enables a reader to identify the interests behind it and to assess its claims. The definition is deliberately confined to what a submitter discloses in the act of lobbying. It excludes the transparency of

the consulting authority, which is a separate construct. Six dimensions are proposed, set out in Table 1. Each can be judged from the submission and its metadata alone.

Table 1: Proposed dimensions of the index, the textual evidence for each and the type of indicator		
Dimension	What is looked for in the submission	Indicator
Identity	Named organisation or person, with legal form or sector	Present or absent
Represented interest	Statement of the members or clients on whose behalf it is made	Ordinal, three levels
Funding	Disclosure of funders, or of who paid for the submission	Present or absent
Register reference	Lobbying or transparency register number, where a register exists	Present or absent
Evidence sourcing	Factual claims attributed to checkable sources	Ordinal, three levels
Coordination	Acknowledgement of shared or template text	Present or absent

3.2. Step 2: Assemble the corpus

The population must be fixed before any text is read: which authority, which consultations, which years. All published submissions to the chosen consultations are collected with their metadata, meaning the date, the consultation identifier, the declared name and category of the submitter and the file itself. Submissions that the authority received and did not publish are recorded as missing, with the reason if it is given, because confidential submissions are exactly the cases a transparency measure should not silently drop. The terms of use of the portal and the applicable data protection law are checked at this point, since submissions from individuals contain personal data.

3.3. Step 3: Clean the texts and link submitters

Submissions arrive as documents of many kinds, some of them scanned. They are converted to plain text, with optical character recognition where needed and a record of which files required it. Language is detected, since multilingual consultations need either translation or separate models. Near-duplicate texts are identified by comparing overlapping word sequences, a standard technique in information retrieval (Manning et al., 2008); this both flags mass campaigns and supplies the factual basis for the coordination dimension. Submitter names are then standardised and linked across consultations and, where a lobbying register exists, to register entries.

3.4. Step 4: Hand-code a sample

A codebook is written that defines each indicator, gives decision rules and supplies examples of borderline cases. Two or more coders apply it independently to a stratified random sample of submissions, drawn so that each submitter category and consultation is represented. Agreement is assessed with a chance-corrected coefficient, such as Cohen's kappa for two coders (Cohen, 1960) or Krippendorff's alpha for more (Krippendorff, 2018). Where agreement on an indicator is poor the codebook is revised and the sample recoded. An indicator on which trained readers cannot agree should be dropped, since no algorithm can recover a distinction that people cannot draw. The reconciled sample becomes the reference set for the next step.

3.5. Step 5: Scale up with dictionaries and a supervised classifier

Some indicators are close to mechanical. A register number has a fixed format and can be found with a pattern search, and identity can often be read from metadata. Others depend on phrasing. For these, two approaches are run side by side. The first is a dictionary of phrases drawn from the coded sample, such as formulations that introduce a membership or a client. The second is a supervised classifier trained on the

reference set, for which a pretrained language model fine-tuned on the coded passages is the current standard approach (Devlin et al., 2019). Part of the reference set is held back from training and used only to measure performance, reported for each indicator as precision and recall. An indicator that the automated methods cannot reproduce acceptably is either hand-coded in full or excluded, and the report says which.

3.6. Step 6: Aggregate to an index

Each submission now has a score on each dimension. Ordinal indicators are rescaled to the unit interval so that all six are comparable. The default proposed here is an unweighted mean, on the reasoning that no dimension has an evident claim to priority and that equal weights are the easiest to explain; the handbook on composite indicators discusses the alternatives and their costs (OECD and JRC, 2008). Dimensions that do not apply, such as a register reference in a jurisdiction with no register, are omitted from the mean and not scored as zero. Submission scores can then be averaged by submitter, by category of submitter, by consultation or by year, with the number of submissions behind each average always shown.

3.7. Step 7: Validate and release

Validation asks whether the scores mean what the definition says. Three checks are proposed. Content validity is examined by asking practitioners and scholars of lobbying whether the dimensions cover the concept. Convergent validity is examined where an independent source exists: organisations that file fuller entries in a lobbying register would be expected to score higher. Robustness is examined by recomputing the index under different weights, with each dimension removed in turn, and with dictionary and classifier scores substituted for one another. A large change in rankings under a small change in method is a finding to report, not to conceal. The codebook, the coded sample, the code and the scores are then released so that others can reproduce and challenge the index (Benoit et al., 2016).

4. Fitting the Procedure to Practice

The procedure can be scaled to the resources available. A single researcher studying one consultation with a few hundred submissions may code everything by hand and omit the fifth step. The automated stage earns its cost only when the corpus is too large to read, and it should not be adopted for its own sake. A team working across many consultations will need a second coder throughout and a modest computing budget for model training.

Jurisdictions differ in ways that affect comparability. Some portals oblige respondents to choose a category and to state a register number, which makes disclosure on those dimensions a matter of compliance and not of choice. Others accept free text with no structure. An index built in one setting will not transfer unchanged to another, and comparisons across jurisdictions should be confined to the dimensions on which the portals leave respondents equal freedom. Where submissions from private individuals are included, names should be removed from any released data.

5. Limitations and What Remains to Be Shown

The procedure is proposed and has not been tested. Whether coders can reach acceptable agreement on the two ordinal dimensions, whether a classifier can reproduce their judgements, and whether the index correlates with independent measures are open questions, and a negative answer to any of them would require the design to change.

Some limits are built in. The index measures disclosure and not truthfulness: a submission that names a funder falsely scores as well as one that names it honestly. It observes only formal, written lobbying through one channel, and organisations that lobby mainly through meetings will be invisible to it. Published consultations over-represent actors with the resources to respond, so averages describe those who took part and not the population of interests. Finally, once an index is known to be in use, submitters may adapt their wording to it, which is welcome if the disclosure is real and a problem if it is formulaic. A measure of this kind should be read alongside register data and case knowledge, never in place of them.

6. Conclusion

Consultation submissions are an abundant and underused record of lobbying, and the tools needed to read them at scale are now ordinary. The procedure proposed here ties those tools to a defined construct, a written codebook, measured coder agreement and explicit validation, so that the resulting number can be defended and reproduced. It would let researchers compare disclosure across types of organisation and over time in settings where no lobbying register exists. Whether it works as intended must be established by applying it, and the article is offered as an invitation to do so.

Conflict of interest

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

Funding

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

References

- Adcock, R. and Collier, D. (2001). [Measurement validity: A shared standard for qualitative and quantitative research](#). *American Political Science Review*, 95(3): 529-546.
<https://doi.org/10.1017/S0003055401003100>
- Baker, S.R., Bloom, N. and Davis, S.J. (2016). [Measuring economic policy uncertainty](#). *Quarterly Journal of Economics*, 131(4): 1593-1636. <https://doi.org/10.1093/qje/qjw024>
- Benoit, K., Conway, D., Lauderdale, B.E., Laver, M. and Mikhaylov, S. (2016). [Crowd-sourced text analysis: Reproducible and agile production of political data](#). *American Political Science Review*, 110(2): 278-295. <https://doi.org/10.1017/S0003055416000058>
- Bertrand, M., Bombardini, M., Fisman, R., Hackinen, B. and Trebbi, F. (2021). [Hall of mirrors: Corporate philanthropy and strategic advocacy](#). *Quarterly Journal of Economics*, 136(4): 2413-2465.
<https://doi.org/10.1093/qje/qjab023>
- Chari, R., Hogan, J. and Murphy, G. (2010). [Regulating Lobbying: A Global Comparison](#). Manchester University Press, Manchester.
- Cohen, J. (1960). [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1): 37-46. <https://doi.org/10.1177/001316446002000104>
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). [BERT: Pre-training of deep bidirectional transformers for language understanding](#). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.
<https://doi.org/10.18653/v1/N19-1423>
- Gentzkow, M., Kelly, B. and Taddy, M. (2019). [Text as data](#). *Journal of Economic Literature*, 57(3): 535-574.
<https://doi.org/10.1257/jel.20181020>
- Grimmer, J. and Stewart, B.M. (2013). [Text as data: The promise and pitfalls of automatic content analysis methods for political texts](#). *Political Analysis*, 21(3): 267-297. <https://doi.org/10.1093/pan/mps028>
- Klüver, H. (2009). [Measuring interest group influence using quantitative text analysis](#). *European Union Politics*, 10(4): 535-549. <https://doi.org/10.1177/1465116509346782>
- Klüver, H. (2013). [Lobbying in the European Union: Interest Groups, Lobbying Coalitions, and Policy Change](#). Oxford University Press, Oxford.
- Krippendorff, K. (2018). [Content Analysis: An Introduction to Its Methodology](#), 4th edition. Sage, Thousand Oaks, CA.
- Loughran, T. and McDonald, B. (2011). [When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks](#). *Journal of Finance*, 66(1): 35-65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- Manning, C.D., Raghavan, P. and Schütze, H. (2008). [Introduction to Information Retrieval](#). Cambridge

University Press, Cambridge.

OECD (2010). [Recommendation of the Council on Principles for Transparency and Integrity in Lobbying](#). Organisation for Economic Co-operation and Development, Paris.

OECD (2021). [Lobbying in the 21st Century: Transparency, Integrity and Access](#). OECD Publishing, Paris.

OECD and JRC (2008). [Handbook on Constructing Composite Indicators: Methodology and User Guide](#). OECD Publishing, Paris.

Yackee, J.W. and Yackee, S.W. (2006). [A bias towards business? Assessing interest group influence on the U.S. bureaucracy](#). *Journal of Politics*, 68(1): 128-139. <https://doi.org/10.1111/j.1468-2508.2006.00375.x>