**Validating measures of ‘information quality’ in stakeholder comments**

**Contents**

[1. Background 1](#_Toc207035454)

[2. What do reviewers demand specifically with regard to validation 3](#_Toc207035455)

[3. What do we want to measure? Relevant excerpts from our manuscript 4](#_Toc207035456)

[4. Draft coding instructions 6](#_Toc207035457)

[5. Options for the outcome variable 7](#_Toc207035458)

[5.1 Single Likert scale 7](#_Toc207035459)

[5.2 Separate Likert scales 8](#_Toc207035460)

[5.3 Pairwise comparisons / BWS 8](#_Toc207035461)

[6. Sampling exemplary texts for coding 10](#_Toc207035462)

## 

## 1. Background

All three *Governance* reviewers asked us to validate our text-based measures for the latent concept of ‘information quality’ in the context of evidence-based policy-making of the European Commission. While the individual comments go into slightly different directions (e.g. whether syntactic complexity or information density qual ‘quality’ in some abstract sense or whether we have the right/valid measures to capture each sub-concept correctly), all of them either explicitly or implicitly require us to develop some kind of human ‘gold standard’ against which we can justify our measurement approach. Coming up with such a comparison set will thus be invaluable to please the reviewers, but will also enhance the contribution (and citeability) of our paper (at the risk of showing that the measures might not work…).

We agreed that the best approach would thus be to gather data on the latent variable ‘information quality’ to then see whether our individual measures are positively (and ideally: strongly) correlated with this latent concept (which also gives us some flexibility as to which measures work best).

I have created such validation tasks in other projects already and can come up with web-based coding applications rather quickly which we can the ship to research assistants to collect their assessments (one example [**here**](https://shiny2.wzb.eu/rauh/EU-Act-0/)).

Two key challenge we face now are (a) formulate respective coding instructions that stay true to what we want to measure in our paper and (b) construct a test sample from our feedback data. This document serves to prepare the respective choices – please edit and comment as you see fit!

## 2. What do reviewers demand specifically with regard to validation

**R1C1**: One major concern is that information quality is directly inferred from the comments’ textual sophistication without any validation. In my view, the contribution of the entire paper hinges on whether we can validly assume that more textually sophisticated comments are also “better” comments with higher information quality. As the paper stands, it is not entirely clear whether the indicators the authors analyze

*-> concept validation needed!*

**R1C4:** Why are high values on word entropy interpreted as indicating high information density? The authors state on page 20 that “the more variable a message’s content is, the denser is the information it contains”. Why? Aren’t messages that spread across multiple concepts and convey a wide range of different pieces of information rather LESS dense than those that are targeted towards a narrower vocabulary? It’s not immediately clear to me why higher entropy should imply higher density. Similarly, why does a more syntactically complex sentence structure imply higher “information quality”? One might argue that when texts become so syntactically complex that they are almost impossible to read, information quality rather declines. One could also argue that the relationship is U-shaped. Again, the main point here is that all of this is assumed, not shown.

* *Clarify in conceptual discussion but also ideally show that higher entropy and complexity is positively correlated with latent ‘information quality’ concept (one could then also test the u-shaped claim)*

**R2C4:** I think the discussion and operationalization of the four dependent variables are great but wondering if the authors can do more to underline their original contribution to the literature in the sense of measuring information quality in all these aspects.

* *Works much better if we can show that the measurement is valid*

**R3C1:** That said, I see one general area where the paper would benefit from substantial expansion: the management of uncertainty and accuracy of the text-as-data measures. The four indicators employed are all plausible proxies that have been used in relevant past literature. However, the manuscript currently treats them as sufficient in themselves, without any empirical check against human judgment. Given that these measures are applied to both policy documents and comments, it is crucial to establish that they align with expert human coders’ assessments of information quality. I would strongly encourage the authors to validate each measure on a sample of documents, ideally 50–100 comments and 50–100 policy texts. Human coders could rate the extent to which each text dimension is present. This would provide a benchmark to demonstrate that the automated indicators capture meaningful variation rather than artifacts of text length, style, or translation. Without such validation, it is not clear whether “syntactic complexity” or “legal terminology” truly maps onto substantive informational quality.

* *Crucial one – while we cannot go for a measure-by-measure validation, showing that all or at least some of them are positively related to the latent concept of information quality should address this comment*

**R3C2:** A related suggestion is to consider incorporating zero-shot large language models (LLMs) as an additional point of comparison. Recent work shows that models such as GPT can perform reasonably well at classification tasks with carefully constructed prompts, without fine-tuning. The authors could, for example, prompt an LLM to classify a random sample of comments along the same four quality dimensions. Comparing those classifications to both human coding and the current automated measures would serve two purposes: (1) it would provide an additional form of validation, and (2) it would illustrate the tradeoffs between traditional supervised NLP, dictionary-based methods, and emerging LLM-based approaches. Even if the LLM approach is not used for the main analysis, reporting the comparison would strengthen confidence in the results and broaden the contribution to the growing text-as-data literature.

* *Not a fan, but once we have human coded data, we can test a zero- or few shot LLM as well (cf.* [*Laurer et al. 2024*](https://www.cambridge.org/core/journals/political-analysis/article/less-annotating-more-classifying-addressing-the-data-scarcity-issue-of-supervised-machine-learning-with-deep-transfer-learning-and-bertnli/05BB05555241762889825B080E097C27)*)*

## 3. What do we want to measure? Relevant excerpts from our manuscript

All quotes come from [GOV\_MS\_consolidated\_v07.03.2025.docx](https://www.dropbox.com/scl/fi/8qd9bcijh499lnsnvcg7e/GOV_MS_consolidated_v07.03.2025.docx?rlkey=1o098i1qdoe3fw5qrtifdh9d2&dl=0)

*‘First, we argue that bureaucratic policymaking is driven by evidence-based decision-making, defined as a process ‘based on the premise that the most adequate and accurate pieces of information are prioritized in decision-making, and that policy effectiveness can be improved through the systematic incorporation of high-quality policy information’ (Migchelbrink et al. 2024, 1). Evidence-based policymaking (EBP) represents a fundamental contextual factor that circumscribes our study of public comments’ information quality insofar as it indicates that this needs to be judged against the standards of evidence-based policymaking (Chalmers 2013; Costa et al. 2019; Yackee 2012). Two implications follow. First, stakeholders recognize that ‘evidence-based policymaking requires evidence-based lobbying’ and policy inputs (Chalmers 2013, 51). Second, if the comments’ information quality is assessed along a continuum, comments richer in evidence and technical details have higher quality than those containing political information in the context of EBP.’*

* Suggests that ‘information quality’ is the key, though latent variable we are interested in

*‘What constitutes evidence in the context of EBP remains an intensely debated topic (Raymaekers et al. 2025). However, there is some agreement that evidence, at the very least, consists of ‘objective information bearing some relationship to a reality that is independent of the observer’ (Jennings and Hall 2012, 246). This information is also analytical as it contains a substantial ‘amount of falsifiable statements about the policy under consideration’, contrasting with ‘non-analytical information – for example conveyed in the form of anecdotes or personal information – which other scholars have categorized as “ordinary knowledge” or “experiential discourse” (Esterling 2007)’ (Ban et al. 2023, 127). Lastly, evidence includes information that is informative in the sense of addressing the policy issues under public consultation (Balla et al. 2024).‘*

* Intermediate clarification of what ‘information quality’ is which could thus be reflected in the instructions (in laymen’s terms)

‘We build on this and argue that in the context of bureaucratic policymaking more generally, and that of the EU supranational policymaking specifically, the information quality of comments is mirrored in different dimensions of their *textual sophistication* (see also Bryer 2013; Shulman 2009). The extent to which comments provide factual/concrete, analytical, and informative information is reflected in the characteristics of the language they use to transmit messages to policymakers. Specifically, and as detailed below, high-quality comments are characterized by (1) high *information density*, (2) analytical arguments demonstrated through *syntactic complexity*, (3) policy orientation indicated by the use of *legal terminology*, and (4) the provision of *factual, concrete, quantifiable* information.’

* The operationalization in the narrow sense and thus the key link to be validated: Do these textual elements link to the latent information quality concept as introduced above

## 4. Draft coding instructions

Here I provide my ideas for the general coding instructions. I have tried to match the conceptual discussion of EBP while making it as concrete for the coders as possible (and not tilting it too much to our measures already). **Please review and edit as you see fit!**

**Coding task: Information quality of stakeholder feedback**

**Your role**: Imagine you are a ***bureaucratic official*** who is tasked with drafting a ***new policy***. You are expected to make sure that this policy reflects ***the best and most accurate information*** available. Thus, you have shared your draft policy with potentially affected societal actors and asked for their ***feedback***.

**Your task**: We show you excerpts from consultation responses. ***Rate how useful each comment excerpt is for helping you make the best factual decision possible***. Ignore whether you agree with the stance or values expressed. Focus only on the **quality of information**.

**Guidance**: High quality information can come in many forms but a few questions may help when rating the exemplary excerpts. ***Relevance*:** Does the comment directly address a regulatory issue or does it rather appear as vague or off-topic? ***Richness***: Does the comment offer much information or is it rather simplistic/limited? ***Analytic content*:** Does the comment provide arguments, factual and checkable claims (e.g., data, studies, mechanisms, legal/technical details), or is it mostly opinion, advocacy, or anecdote? ***Specificity***: Is the information concrete and detailed enough to be directly useful for legal policymaking, or is it rather general and unspecific?

## 5. Options for the outcome variable

I see the different options for measuring the outcome after coders have seen the general coding instructions above (those remain on the screen, while example texts switch after each click. Each comes with advantages and disadvantages that I want to quickly lay open here.

### 5.1 Single Likert scale

Here the coders would see one excerpt at a time to rate its overall information quality on an absolute scale (Likert), e.g. (you may edit here as well):

*From the contents of the new delegated act it emerges that the possibility of reporting personnel expenses in the context of operational programs would be limited only to the members of pos with the legal form of a cooperative company, thus excluding other corporate forms and introducing a limitation which would in fact lead to discriminate support not on the basis of the activity carried out but on the basis of the person who carries it out. In italy, the typology of po members includes, in addition to cooperatives, other corporate forms such as consortium companies or limited liability companies to which many producers belong and which create a effective concentration of large volumes of fruit and vegetables.*

***Rate the overall information quality of this response:***

**1 = Very Low:** Off-topic or purely opinion/political statement; no usable evidence.

**2 = Low:** On-topic but anecdotal, vague, or mostly experiential without verifiable support.

**3 = Moderate:** Some relevant facts or analytic points, but limited detail, sourcing, or clarity.

**4 = High:** Mostly relevant and evidence-based; includes concrete data, examples, or analysis that could inform a decision.

**5 = Very High:** Directly relevant, well-supported, specific evidence or analysis; clearly useful for evidence-based policymaking.

**Pros**: Super easy to implement in the app, correlation with our text-based measures relatively straightforward

**Cons:** Rating on an absolute scale is super-hard, especially for coders inexperienced with EU lingo. Their internal benchmarks of what is useful or not might change significantly over the course of the task (the more they have seen), leading to instable measurement

### 5.2 Separate Likert scales

Here we would ask coders directly to assess sub-dimensions of information quality directly, for example along the guiding questions above:

***Relevance*:** □ 1, □ 2, □ 3, □ 4, □ 5

***Richness***: □ 1, □ 2, □ 3, □ 4, □ 5

***Analytic content*:** □ 1, □ 2, □ 3, □ 4, □ 5

***Specificity***: □ 1, □ 2, □ 3, □ 4, □ 5

**Pros:**

* Significantly easier for coders
* Closer to what R3 demands

**Cons:**

* Same problem as above, absolute benchmark may change over the course of the coding task
* Probably very skewed scales, with rare values in the extremes
* Would require a careful conceptual distinction of the subdimensions, assumes that they are independent, no validation with regard to the overall idea of ‘information quality’

AB thoughts:

Pros also include lower setup costs so somewhat less work for Christian. BUT depends a bit more on the quality of coders and their work. Saving time and work for Christian but depending more on coders and their timelines.

Cons 2: Could we use a 3-point scale (low, medium, high) instead of a 5-point scale?

Cons 3: Could also ask them to assess the overall information quality of the comment and then the specific dimensions? Although I am fairly certain that this introduces some form of bias into our coders’ assessments and scores and all methodologists in the world would advise against it. ☺

***Overall information quality of comment:*** □ 1, □ 2, □ 3, □ 4, □ 5

***Relevance*:** □ 1, □ 2, □ 3, □ 4, □ 5

***Richness***: □ 1, □ 2, □ 3, □ 4, □ 5

***Analytic content*:** □ 1, □ 2, □ 3, □ 4, □ 5

***Specificity***: □ 1, □ 2, □ 3, □ 4, □ 5

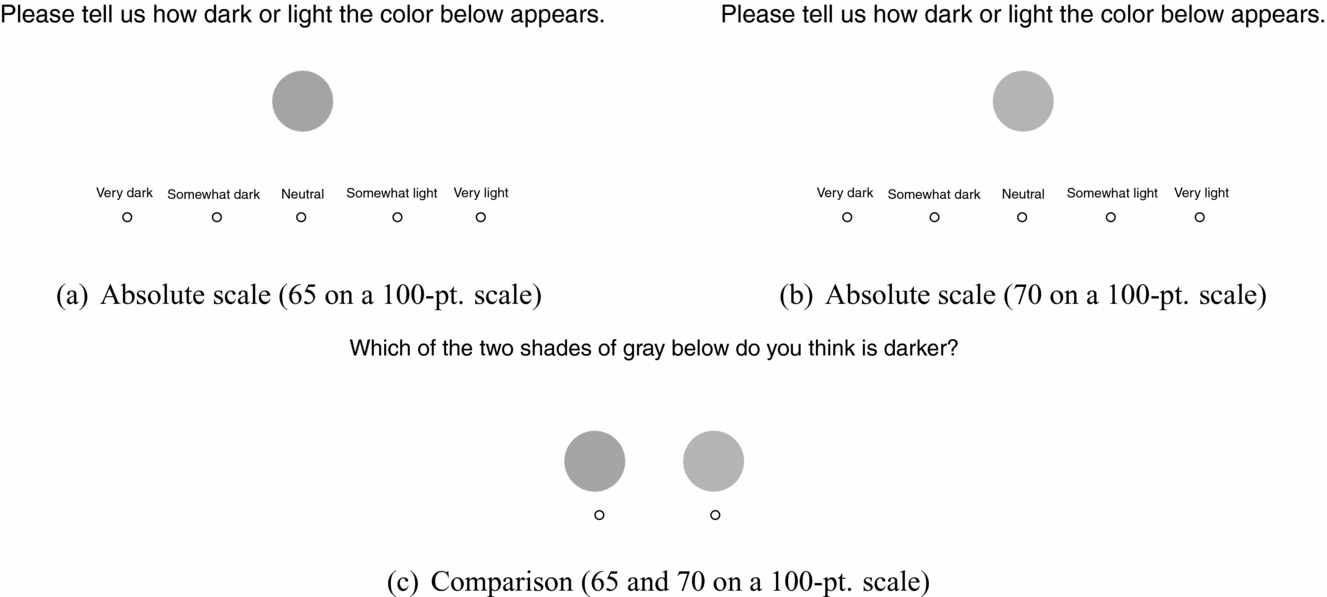
### 5.3 Pairwise comparisons / BWS

In this final approach (which would be quite state-of-the-art), coders would see pairs of feedback excerpts and would rate them in relative terms, e.g.:

|  |  |
| --- | --- |
| **Feedback A** | **Feedback B** |
| i have the following suggestions: art 2(4) please synchronize definition of “validation” with that one given in eu gmp annex 15 (e.g. validation is normally not related to equipment - which is subject to qualification - but to processes) art 9 whereas in art 10 (quality control) the head of quality control is addressed, the head of manufacture with his qualifications is not mentioned in art 9. Please amend accordingly. | From the contents of the new delegated act it emerges that the possibility of reporting personnel expenses in the context of operational programs would be limited only to the members of pos with the legal form of a cooperative company, thus excluding other corporate forms and introducing a limitation which would in fact lead to discriminate support not on the basis of the activity carried out but on the basis of the person who carries it out. In italy, the typology of po members includes, in addition to cooperatives, other corporate forms such as consortium companies or limited liability companies to which many producers belong and which create a effective concentration of large volumes of fruit and vegetables. |
| □ | □ |
| ***Which comment has higher information quality in your view?*** | |

**Pros:**

* For latent variables, relative comparison if often much easier and much more reliable for coders tasked with identifying an abstract latent concept – this can be easily seen along this illustration 😉

 ([source](https://www.cambridge.org/core/journals/american-political-science-review/article/pairwise-comparison-framework-for-fast-flexible-and-reliable-human-coding-of-political-texts/017BF6B024228962FDF90B47FD90EF5F))

* Implicit coder bias (internal benchmarks of information quality) is not such a big issue here
* Based on pairwise comparisons (or best-worst scaling in small groups of four or so) models exist to extract the relative value of a latent concept (e.g. [here](https://www.cambridge.org/core/journals/american-political-science-review/article/pairwise-comparison-framework-for-fast-flexible-and-reliable-human-coding-of-political-texts/017BF6B024228962FDF90B47FD90EF5F), [here](https://onlinelibrary.wiley.com/doi/10.1111/ajps.12423), or here). For validation, we could then assess our measures based on how well the predict (or correlate with) this latent variable
* State of the art approach

**Cons:**

* This would involve quite some programming work for the coding app
* I would need to dig into sampling (see also below) as one needs to make sure that each individually comment is compared sufficiently often to other examples.
* Statistical models become somewhat more complex

AB thoughts:

This approach depends less on the quality of coders and their timelines but implies more work and time for Christian.

R3 might not be entirely happy. We validate the concept, but not individual dimensions. Although we can point out to him that the validation tests he asks for are more pertinent for a methodological paper and not a substantive one. See my last note below on this issue.

## 6. Sampling exemplary texts for coding

The final crucial question is what the coders get to see. Two general challenges occur here in my view.

First, the feedback texts are often way too long for human coders to digest (especially as we want to compare across a larger sample to get meaningful validity assessment). Here I would propose to pick then a fixed set of sentences (2-4, to be determined by inspecting examples qualitatively) from the beginning of each feedback text. OK?

AB: I agree that reading the entirety of a comment might be too much. I would go for the first paragraph of each comment. Not the first sentences only because these might include the names of the organizational stakeholders. Oftentimes, organizations use the first sentence to introduce themselves as organizational actors. Of course information in and of itself says something about the information quality of the comment.

Second, the validation sample should ensure sufficient variation in our text-based measures. I thus would recalculate our text-based measures for the shorter snippets, too then draw a stratified random sample. OK?

AB: OK.

If we go for a pairwise comparison (along option 5.3 above), then it is probably reasonable that coders only compare feedback to the same consultation event, to keep the policy area stable and avoid potential biases in this regard. Combining this with the need to draw a stratified sample, our sample would probably contain more texts from the events that have received more responses in total. Do you see a problem here?

AB: If more comments indicate higher stakeholder salience of a consultation event/act type, then maybe we are validating more our measures for consultation events/acts that are relevant for stakeholders. But I am ok with that!

|  |  |  |  |
| --- | --- | --- | --- |
| Option 5.2 | | Option 5.3 | |
| Costs | Benefits | Costs | Benefits |
| Unsure whether we can validate the latent concept which may pose challenges in answering R1C1, R2C4 | Covers all reviewers’ comments related to the validation issue | Especially time-consuming and labour intensive for CR | Easier and thus perhaps faster for coders to handle. |
| Increases reliance on human coders and their timelines |  | Does not cover directly R3C1 | Covers R1C1, R1C4, R2C4 |
|  |  |  |  |
| AB: If we find good coders that can work on our timeline, I would prefer option 2 as long as we find a way to also validate the latent concept and Christian is comfortable with this less state of the art approach. The advantage of this approach is that it covers more directly R3’s request for measure-by-measure validation and reduces somewhat Christian’s workload.  But if finding good coders proves difficult, or if we want for our timeline to depend less on coders and more on us, then option 3 is a very good option. But it implies a lot more work for Christian. | | | |

NOTE: Some of the validation tests required by R3 (for example, in R3C2) go beyond the scope of a substantive paper and perhaps are more suitable for a methodological paper on the topic. Maybe it is a good idea to recognize this upfront in our response to reviewers’ comments.