Online Appendix

This Appendix illustrates in detail the decisions adopted to apply the Maricut-Akbik (2021) "Q&A approach to legislative oversight" to the analysis of the Recovery and Resilience Dialogues (RRDs). To fully understand the content and the rationale of the changes, our advice is to read this Appendix in conjunction with the Maricut-Akbik (2021) Online Appendix.

1. Variables

The specific features of RRDs led us to adopt various modifications to the framework by Maricut-Akbik (2021). Firstly, follow-up questions are not present, because the procedural rules of the RRDs do not allow follow-ups from MEPs. Secondly, the formal role of the questions' addressees is not assessed, because in every RRD the questions have been posed to the same people: the Executive Vice President of the European Commission (Valdis Dombrovskis) and the European Commissioner for the Economy (Paolo Gentiloni).

In our analysis, we add a variable on the 'level' of accountability: this categorical variable describes whether the topic of the question focuses on the EU level, the national level (when an MEP asks a question about his/her own country), or the transnational level (when an MEP asks a question about a different country from his/her own).

We add a new type of oversight question, "Irrelevant question", in addition to Demand information, Justification of conduct, Change of conduct, Sanctions for actors, and Request policy views. "Irrelevant questions" refer to issues that are outside the policy scope for which the RRD has been established in the first place (Article 26(1)(a-g) RRF Regulation). For instance, a question about the Commission's negotiation of the OECD Global Minimum Tax cannot find any basis in the RRF Regulation.

As regards the variable capturing the answers' content (which are not used in the article in *Politics and Governance*) we have added the category "Evasion" to distinguish it from the category of "equivocation", which we felt was too encompassing in the original Q&A approach. In our framework, "Equivocation" answers are the ones where the addressee engaged with the question, while at the same time not understanding (or pretending to not understand) it; "Evasion" indicates that the question was completely ignored.

Finally, we have added the categories "Partial answer to change request", and "Partial answer to sanctions request", and allowed the use of "Provide policy views" and "Answer to irrelevant questions" as content of intermediate replies. In this way, we can describe cases where the addressees do not give complete or exhaustive answers, that cannot be labelled with the categories of "Invoke secrecy", "Promise future information", or "Different actor's responsibility".

Therefore, the variables that we coded are presented in Table A, and the resulting codebook is summarised in Table B.

Table A: Variables included in the dataset

Question	Claim aspect	Variable name(s)
When was the claim made?	• Date of the RRD	• date
Who is making the claim?	• Claimant committee (BUDG/ECON/BUDG & ECON) • Claimant nationality • Claimant EP group affiliation	• committee • country • affiliation
What is the claimant demanding?	• Type of oversight question (demand information/justification of conduct/change of conduct/sanctions for actors/policy views/irrelevant question) • Accountability focus (EU/national/transnational/)*	• oversight_question • level*
HOW is the addressee responding? (degree)	• Type of answer (explicit reply/ intermediate reply/non-reply) • Content of answer (Transparency/ Justification/Accept change/defend decision/Accept sanctions/defend conduct/Sanctions no longer necessary/Provide policy views/Answer to irrelevant questions/Invoke secrecy/Promise future information/Different actor's responsibility/Equivocation/Evasion)	• answer_type • answer_content

**not present in Maricut-Akbik (2021).*

Table B: Relationship between the type of questions, the type of answers and the content of the answer

	Explicit reply	Intermediate reply	Non-reply
Request information	Transparency	Transparency* Invoke secrecy Promise future information Different actor's responsibility	Evasion* Equivocation Invoke secrecy
Request justification	Justification	Justification Invoke secrecy Promise future information Different actor's responsibility	Evasion* Equivocation Invoke secrecy
Request a change	Accept change Defend decision	Partial answer to change request* Invoke secrecy Promise future information Different actor's responsibility	Evasion* Equivocation Invoke secrecy
Request sanctions	Accept sanctions Defend conduct Sanctions no longer necessary	Partial answer to sanctions request* Invoke secrecy Promise future information Different actor's responsibility	Evasion* Equivocation Invoke secrecy
Request policy views	Provide policy views	Provide policy views* Invoke secrecy Promise future information Different actor's responsibility	Evasion* Equivocation Invoke secrecy
Irrelevant question*	Answer to irrelevant question*	Answer to irrelevant question* Invoke secrecy* Promise future information* Different actor's responsibility*	Evasion* Equivocation* Invoke secrecy*

** not present in Maricut-Akbik (2021).*

2. Intercoder reliability check

To pursue an intercoder reliability test, we recoded a sample of about 40% of the full dataset (132 claims out of a total of 335). To make the procedure consistent with the first coding process, we decided to recode entire interventions rather than individual claims. This helps the coder to contextualise the questions in the whole interventions of the MEP.

The procedure to create the stratified sample was the following. The interventions originally coded by CoderA, CoderB and CoderC are kept in separate files. A random sample of interventions is selected from the ones coded by CoderA. If the interventions are all coming from only one RRD, the random selection is repeated. Then, the interventions are stripped out of their content (only data on the name of the MEP and the date of the RRD remain), and half of them are assigned to CoderB, and half are assigned to CoderC, so they can independently recode the interventions. The procedure is then repeated for the interventions of CoderB and CoderC. Overall, each coder has received about 44 claims previously coded by the two other coders.

After the process of recoding had been completed, three indices were used to measure inter-coder reliability:

- **Percent agreement:** it is the percentage of decisions, made by pairs of independent coders, on which the coders agree. It is the most simple and intuitive index to calculate. However, the current literature contests the exclusive use of this measurement to calculate inter-coder reliability, because it does not take into account the agreement that occurs simply by chance (Lombard et al. 2002).

- **Krippendorff's alpha (α):** it is an index that accounts for chance agreement, by comparing the actual agreement of coders with the percent agreement the reviewers would achieve by randomly guessing. Its attractiveness lies in the fact that it can handle missing data and is comparable across any number of coders, types of metrics (nominal, ordinal, ratio, etc.), and unequal sample sizes (Krippendorff, 2004).

- **Gwet AC:** this index also takes into account chance agreement, but by comparing the actual agreement of coders with the percent *disagreement* the reviewers would achieve by randomly guessing. It can handle the same data type as Krippendorff's alpha and is increasingly considered as a more stable agreement coefficient because it is less sensitive to the prevalence of certain traits in the analysed population (Gwet 2008, see further below).

When calculating the Krippendorff's alpha and Gwet AC scores of the variables "Question type" and "Answer type", we consider these variables as ordinal, thus applying a weight matrix (therefore, we use the weighted version of Gwet AC, that is AC_2). "Question type" measures the strength of accountability of the question, with the following order from the lowest to the strongest value: Irrelevant question, Request policy views, Request information, Request justification, Request a change, Request sanctions. "Answer type" measures the exhaustiveness of the Commissioners' answers, with the following order from the lowest to highest value: Non-reply, Intermediate reply, Explicit reply.

There is no consensus on the minimum values certifying the reliability of coding. Landis and Koch (1977) provide guidelines for Cohen's Kappa with values from 0.0 to 0.2 showing slight agreement, 0.21 to 0.40 as fair agreement, 0.41 to 0.60 as moderate agreement, 0.61 to 0.80 as substantial agreement, and 0.81 to 1.0 as almost perfect agreement. Krippendorff (2004) is more conservative and claims that values lower than 0.67 should be considered as unreliable, while tentative conclusions can be made for values between 0.67 and 0.80, and definite conclusions can be made for values above 0.80.

Table C shows the results of the inter-coder reliability test. By using the thresholds of Krippendorff (2004) and looking at Table C, the coding can be considered sufficiently reliable for each of the analysed variables and indexes. The only exception is the Krippendorff's alpha score for the Answer type, which scores below 0.67 despite having a high percent agreement.

Table C. Percentage agreement of the variables analysed in the paper.

Variable	**Percent agreement**	**Krippendorff-alpha**	**Gwet AC_2**
Question type	0.7348	0.7054	0.9430
Answer type	0.7727	0.5870	0.7810
Focus	0.9242	0.7820	0.9345

We hypothesize that such a score might be partially due to the "paradoxical" behaviour of Krippendorff's alpha. When using this index (and other indices calculating chance agreement in a similar way, such as Cohen's Kappa and Scott's Pi), if the prevalence of a certain trait is low or high in the population, a large extent of agreement between raters will not be reflected in the final score (Feinstein and Cicchetti, 1990, Gwet 2008). This is exactly the case for the Answer type variable: the population number of Explicit replies (n=226) is significantly higher than the number of Intermediate replies (n=69) and Non-replies (n=37). The influence of this element, combined with a solid Gwet AC_2 score for Answer type, allows us to consider the coding of that variable as robust.

3. The coders' *vademecum*

Along the process of discussion of the framework, we have established a series of practical coding rules (a '*vademecum*') to obtain the highest intercoder consistency. We are fully aware that different coders and contexts might produce different decisions (and, therefore, slightly different variables). Our aim was to enforce a consistent interpretation of the codebook, not to reach any kind of 'objectivity', which is impossible given the highly interpretative nature of coding political speech acts. The *vademecum* is the following:

- "Request information" questions have a neutral tone and mere descriptive goals. Such questions are about past conduct and decisions that have already been taken, or actions that it is reasonable to think have already been taken (executive or implementing procedures).
- "Request justification" questions must: (i) be about conducts or decisions that have already been taken; (ii) be morally/value charged; and (iii) refer – explicitly or implicitly – to a possible alternative reality.
- "Request a change" questions must include an *explicit* proposal to act on a decision that has already been taken.
- "Request policy views" questions are recognisable because they (i) do not focus on *past* behaviour of the Commission; contrarily, their point to *future* choices and decisions; (ii) do not target the Commission's executive power on the RRF, instead inquiring the Commission in its role of legislator, agenda setter, etc.
- On the difference between "Request policy views" and "Irrelevant questions": the latter refer to issues that are outside the policy issues for which the RRD has been established, as indicated by Article 26(1)(a-g) of the RRF Regulation; the former request speculative views, opinions and considerations on future actions (Maricut-Akbik 2021, Appendix, p. 22) that are explicitly *linked* to the policy issues contained in the RRDs' scope. This link can be merely mentioned by the MEP.
- If an MEP mentions both his/her own country and another country in the question, the level is coded as "transnational".

References

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*, 43(6), 543-549.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1), 29-48.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, California: Sage.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human communication research*, 28(4), 587-604.

Maricut-Akbik, A. (2021). Q&A in Legislative Oversight: A Framework for Analysis. *European Journal of Political Research*, 60(3), 539–59.