

Probabilistic Narrative Identification: Evidence from Tax Policy

Online Appendix

Tom Pessó

October 12, 2026

Contents

OA.1	Measurement Documentation	3
OA.1.1	From Policy Documents to Narrative Support	3
OA.1.1.1	Embedding Models and Distance Metric	3
OA.1.1.2	Retrieval Performance across Parameter Settings	4
OA.1.1.3	Motivation-Extraction Prompt and Settings	4
OA.1.1.4	Example details	4
OA.1.1.5	Grounding Diagnostics: Marginal Narrative Contributions	6
	Global Placebo Test for Summarisation Grounding	6
	Placebo mass.	6
	Randomisation inference.	6
	Interpretation.	6
	French VAT results.	7
OA.1.2	Taxonomy and Probability Construction	7
OA.1.2.1	Classification Prompts and Category Scheme	7
OA.1.2.2	Synthetic Calibration Design	16
OA.1.2.3	Probability-Map Estimation	17
OA.1.2.4	Model Replication Details	17
OA.1.2.5	Alternative LLM Signal: Incremental Log Probabilities	17
OA.2	Supplementary Calibration Diagnostics	19
OA.2.1	Synthetic construction diagnostics	19
OA.2.2	Dimensionality of the Hidden-State Representation	19
OA.3	Supplementary Romer–Romer Evidence	22
OA.3.1	Probability Benchmark Settings	22
OA.3.1.1	Narrative extraction	22
OA.3.1.2	Overlap Between the RR Series and Budget-Based Measures	24
OA.3.2	Estimator-Specific Romer–Romer Distributed-Lag Diagnostic	26
OA.3.3	Two-type endpoint diagnostic	26
OA.3.4	Narrative Ambiguity in the Romer–Romer Probability Benchmark	26
	Clean alignment.	28
	Boundary exogenous cases.	29
	Deficit and Social Security cases.	29
	Departures that reveal category boundaries.	30
	Takeaway.	30
OA.3.5	Training-data leakage robustness check	30
	Procedure.	30

	Decision rule.	31
OA.3.6	Differences Relative to the Romer–Romer Classification	32
	Steering growth above normal levels.	32
	Conflicting motivations.	32
	“Package” related measure.	32
OA.4	Additional U.S. Composition and Transmission Evidence	33
OA.4.1	Affine GMM Endpoint Diagnostic	33
OA.4.2	Component Diagnostics and Romer–Romer Comparison	33
OA.4.3	Complete Instrument-Specific Response Paths	36
OA.5	A Structural Tax-Wedge Benchmark	39
	Structural benchmark and pure tax wedges.	39
	Mapping the empirical income-tax category into model wedges.	40
	Composition sensitivity.	40
	Consumption taxes across calibrations.	42
	Empirical comparison in GDP units.	42
	Robustness and limitations.	42
OA.6	Supplementary VAT Diagnostics	44
OA.6.1	Additional VAT Channel Diagnostics	44
OA.6.2	Complete Paths Across Support Constructions	46
OA.6.3	Complete Package-Sensitivity Paths	46
OA.6.4	Country-Specific VAT Diagnostics	46

OA.1 Measurement Documentation

OA.1.1 From Policy Documents to Narrative Support

This appendix complements the conceptual measurement discussion in Paper Section 2 and the implementation choices in Paper Section 3. It records the document mapping, application-specific retrieval choices, motivation-extraction settings, and grounding diagnostics used to trace each support score back to its source material.

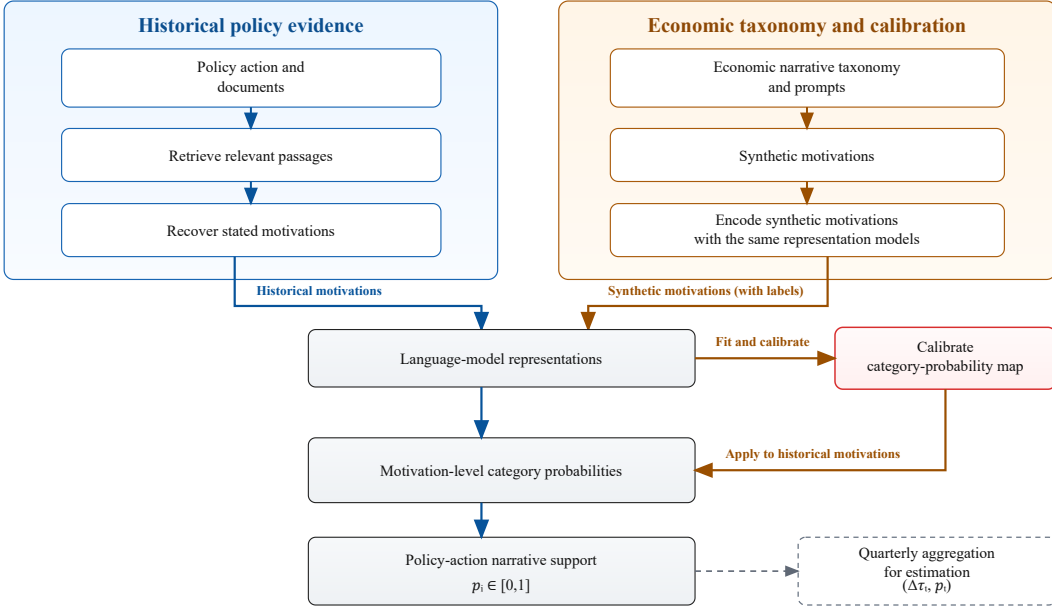


Figure OA.1: From documentary evidence and an economic taxonomy to narrative support. The historical branch recovers stated motivations from policy-specific documents. The calibration branch uses an economic taxonomy and category prompts to construct labeled synthetic motivations and estimate a category-probability map. Both sets of motivations are encoded using the same language-model representations. Applying the calibrated map to the historical motivations yields motivation-level category probabilities, which are aggregated into policy-action narrative support p_i . Quarterly aggregation produces $(\Delta\tau_t, p_t)$ for estimation.

OA.1.1.1 Embedding Models and Distance Metric

Semantic retrieval relies on vector representations of text segments and policy descriptions. The embedding model differs across applications. Table OA.1 records the models and retrieval depth used in each case.

Table OA.1: Application-Specific Retrieval Settings

Application	Embedding model	Retrieved passages K
French VAT	OpenAI <code>text-embedding-3-large</code>	10
Cloyne validation	OpenAI <code>text-embedding-3-small</code>	10
U.S. budget documents	Google <code>text-embedding-004</code>	10
Romer–Romer validation	Not applicable	Not applicable

Note: The Romer–Romer exercise starts from edited narrative entries and therefore does not require document retrieval.

For each policy description q_i and document segment $d_{i,k} \in \mathcal{D}_i$, the corresponding embedding vectors e_i and $v_{i,k}$ are normalized to unit length. Semantic proximity is then evaluated using the cosine distance,

$$\text{dist}_{\cos}(q_i, d_{i,k}) = 1 - \frac{e_i \cdot v_{i,k}}{\|e_i\| \|v_{i,k}\|},$$

so that smaller distances indicate greater semantic similarity. Because all embeddings are normalized to unit length, cosine and Euclidean distances are mathematically equivalent in this setting, yielding identical retrieval rankings under either metric. Distances are computed through standard nearest-neighbor search routines provided by the underlying vector store, which efficiently identifies the text segments closest to each policy query in the embedding space. This normalization ensures that retrieval reflects semantic similarity rather than differences in vector magnitude. The retrieval operator defined in Equation (6) in the main text therefore selects, for each policy event i , the K document segments with the smallest cosine distances to the corresponding policy query q_i .

OA.1.1.2 Retrieval Performance across Parameter Settings

I use the French VAT corpus to assess retrieval quality in the fiscal-policy setting. Each of the 94 VAT measures is manually linked to the Senate-report article discussing the corresponding budget-law provision. For each candidate chunk length, I record whether the retrieved passages overlap the manually matched article. Table OA.2 reports top-1 retrieval, the share of measures with at least one match among the top three and top ten passages, and the average number of matches among the top ten.

Table OA.2: French VAT Retrieval Performance across Chunk Lengths

Chunk length (characters)	Top-1 retrieval	Top-3 recall	Top-10 recall	Mean matches in top 10
512	0.915	0.968	1.000	6.70
1,024	0.936	0.989	1.000	6.14
2,048	0.894	0.947	0.968	5.19
4,096	0.915	0.979	1.000	4.98

Note: The sample contains 94 French VAT measures. Each measure is matched manually to the relevant article in the corresponding Senate report. Retrieval uses the policy-description query employed in the baseline construction and OpenAI’s *text-embedding-3-large* model.

Performance is high across the parameter grid. The correct article appears among the ten highest-ranked passages for between 96.8 and 100 percent of measures, while top-1 retrieval ranges from 89.4 to 93.6 percent. Although the French Senate reports differ in structure from the UK and U.S. source collections, this benchmark suggests that retrieval is not sensitive to the chunk lengths considered.

OA.1.1.3 Motivation-Extraction Prompt and Settings

The extraction prompt asks why the tax measure was adopted rather than what the measure did. It instructs the model to rely only on the retrieved passages, ignore unrelated policies, avoid external interpretation, and return one to three ordered motivations without introductory text. The main estimates condition on a fixed set of extracted motivations. For the U.S. application, the model used to extract the baseline motivations was not recorded; the extracted texts are included in the replication materials, so the subsequent probability construction and estimation are reproducible conditional on these motivations. Table OA.3 records the settings used in the separate open-weight extraction sensitivity exercise.

OA.1.1.4 Example details

Applying the retrieval operator $\mathcal{R}_K$ to the query q_i and corpus $\mathcal{D}_i$, I obtain the top- K most relevant passages. Here are the top-5 passages:

Table OA.3: Motivation-Extraction Settings

Setting	Sensitivity design
Open-weight models	Qwen-9B, Gemma-4B, and Mistral-7B
Maximum retrieved passages	10
Maximum generated tokens	512
Temperature	0.1
Seeds and draws	Twenty fixed-seed draws per model
Output format	One to three ordered motivation bullets

- **Chunk 1.** On December 31, 1976, to mitigate, by a reduction of the weight of taxes bearing on consumption, the upward movement likely to occur at the end of this phase of price freeze; — it constitutes the last stage of a reform undertaken since 1969 — the date on which the standard rate of VAT was 23.456% — which results, ultimately, in a reduction, in less than seven years, of one quarter of the standard rate and in the merging of the latter with the intermediate rate.
- **Chunk 2.** An exceptional contribution would be requested in 1977 from the wealthiest French; it would affect all of them and only them, contrary to what had been provided for in the deleted article 5. This taxation would have as its base the total of the bases corresponding to the elements of lifestyle enumerated in article 168 of the General Tax Code, with the exception of main residences and cars of a power equal to or less than 16 horsepower. The income-equivalent threshold, calculated according to the norms of the aforementioned article 168, above which the exceptional contribution applies, is set at 60,000 F and the number of accumulated elements provoking taxation is three. Your Finance Committee proposes that you adopt the present article. VAT (Value Added Tax) – Article 6. Reduction of the standard rate and unification of the standard rate and the intermediate rate. Text. — I. — The standard rate of value-added tax is set at 17.60%. In the departments of Guadeloupe, Martinique and Réunion, it is set at 7.50II. — The upper limit of the special rebate for artisans fixed in paragraph 3 of article 282 of the General Tax Code is raised to 20,000 F. III. — The provisions of the present article enter into force on January 1, 1977. Comments. — It is proposed in the present article to lower by 2.4 points, starting from January 1, 1977, the standard rate of value-added tax which is currently 20% and, thereby, to bring it down to the level of the intermediate rate, namely 17.60%. This measure fits within a double perspective: — it is part of the program of combating inflation and is of a nature to permit, after the price freeze applied during the period from September 15 to December 31, 1976, to mitigate, by a reduction of the weight of taxes bearing on consumption, the upward movement likely to occur at the end of this phase of price freeze;
- **Chunk 3.** The number of rates applied in the matter of value-added tax, starting from next January 1, will be reduced to three instead of four; there would remain only: — the increased rate (33 %: cars, motorcycles, photography and cinematography equipment, sound and image recording devices, precious stones and precious metals, etc.); — the standard rate (17.6%: energy products, beverages, services and all transactions hitherto subject to the 20% rate, that is to say all those which were not taxed at the increased rate, nor at the intermediate rate, nor at the reduced rate); — the reduced rate (7%: essentially food products, pharmaceutical products, water and fertilizers). At the same time, it is planned to carry out a reduction of 2.5 points of the standard VAT rate and to align it with the intermediate rate in the departments of Guadeloupe, Martinique and Réunion, French Guiana not being concerned since the value-added tax regime is not applied there. In the three departments concerned, where, due to constraints inherent in the economies of developing countries, VAT rates are significantly lower than those of Metropolitan France, starting from January 1, 1977, one will record: — the increased rate: 14%; — the standard rate: 7.5%; — the reduced rate: 3.5%. In order to support the artisanal professions, it is proposed in paragraph II of the present article to raise, starting from January 1, 1977, the

upper limit of the special rebate from which they benefit, raising it from 13,500 F to 20,000 F. Let us recall that, in order to lighten the burdens of small taxpayers, and in particular artisans, article 282 of the General Tax Code has provided, in their favor, special provisions resulting in a waiver or a reduction of the VAT normally incumbent upon them.

- **Chunk 4.** Standard of 6.5% per year. b) The elimination of the levy. As had been provided by the Finance Law for 1976, in the initial text presented by the Government, the levy was to cease to apply when the progression of prices was below a set norm: this one, in the governmental project, was fixed at 2.50% for the consumer price index of the group of manufactured products during a period of six consecutive months.
- **Chunk 5.** Article: — in paragraph I, the objective is to encourage business leaders to convert into capital the advances they are led to grant in the form of deposits in current accounts to the companies they manage. Up to now, in such cases, the ordinary contribution right was a proportional duty of 1%; it is planned to substitute for it a fixed duty of 220 F on the condition, however, that the sums thus incorporated into the capital have been placed at the constant disposal of the company, for a minimum period of twelve months.

OA.1.1.5 Grounding Diagnostics: Marginal Narrative Contributions

Global Placebo Test for Summarisation Grounding The coefficient-level placebo comparison in the main text evaluates, for each passage individually, whether its positive marginal contribution is unusually large relative to the distribution of positive placebo contributions. Because this diagnostic is applied to multiple passages within each event, some flagged cells may be mechanically expected under the null of no grounding. This appendix defines an omnibus diagnostic that aggregates evidence across passages.

Placebo mass. For each event i , let $w_{i,k}^+ = \max\{\hat{w}_{i,k}, 0\}$ denote the positive marginal contribution of passage k , and let $\mathcal{I}_i^{\text{rel}}$ and $\mathcal{I}_i^{\text{pl}}$ denote the sets of relevant and placebo passages, respectively. The share of positive weight falling on placebo passages is

$$\text{PM}_i = \frac{\sum_{k \in \mathcal{I}_i^{\text{pl}}} w_{i,k}^+}{\sum_{k \in \mathcal{I}_i^{\text{rel}} \cup \mathcal{I}_i^{\text{pl}}} w_{i,k}^+} \in [0, 1],$$

which summarises, at the event level, how much of the model’s total positive dependence is placed on semantically irrelevant text. Grounded behaviour is characterised by low PM_i , while ungrounded dependence yields large values.

Randomisation inference. To assess whether PM_i is unusually low, I test the sharp null that placebo labels are exchangeable with relevant-passage labels conditional on the estimated contributions. Under the null, any assignment of $|\mathcal{I}_i^{\text{pl}}|$ placebo labels among the $|\mathcal{I}_i^{\text{rel}}| + |\mathcal{I}_i^{\text{pl}}|$ passages is equally likely. I therefore permute these labels while preserving the number of placebo passages, recompute the placebo mass $\text{PM}_{i,b}$ for each permutation, and form the lower-tailed randomization p -value

$$p_i^{\text{rand}} = \frac{1 + \sum_{b=1}^B \mathbb{1}\{\text{PM}_{i,b} \leq \text{PM}_i\}}{1 + B}.$$

A small p_i^{rand} indicates that positive dependence is unusually concentrated on the retrieved policy-relevant passages. Failure to reject does not establish that the motivation is ungrounded; it means that the attribution pattern does not distinguish relevant from placebo passages under this diagnostic.

Interpretation. This global diagnostic complements the passage-level exercise in the main text. The passage-level comparison identifies which retrieved passages receive unusually large positive contributions, whereas the global statistic asks whether positive dependence is concentrated on relevant passages at the event level.

French VAT results. For the French VAT corpus, complete attribution outputs are available for 400 motivations associated with 140 policy actions. Averaging the OLS attribution coefficients across the Qwen and Mistral evaluators, the global test rejects exchangeability between retrieved and placebo passages for 70.5 percent of motivations. The median motivation assigns only 30.4 percent of its positive attribution weight to placebo passages, substantially below the 75 percent expected if attribution were unrelated to passage relevance. At the policy level, 132 of 140 actions contain at least one motivation that passes the test; only eight contain none. This exercise covers French VAT proposals, a subset of the tax changes studied in the paper.

OA.1.2 Taxonomy and Probability Construction

OA.1.2.1 Classification Prompts and Category Scheme

This appendix reports the prompt templates that define the eight narrative categories used in the tax applications. Each category corresponds to one of the subtypes summarized in Paper Figure 2, following the narrative taxonomy of Cloyne (2013). The prompts should be read as the semantic dictionary of the measurement problem, not as the main probability estimator. In the baseline construction, these definitions guide the generation of labeled synthetic motivations, and hidden-state features are calibrated on that synthetic corpus. In the alternative signal in Appendix OA.1.2.5, the same prompts are used directly to compute the model’s probability of answering *Yes*.

Prompt templates

Each prompt is presented in a standardized format and corresponds to one category listed in Paper Figure 2. The prompts are organized by group: *endogenous* (Demand Management OA.1, Supply Stimulus OA.2, Deficit Reduction OA.3, Spending-driven OA.4) and *exogenous* (Long-term OA.5, Ideological OA.6, External OA.7, Deficit Consolidation OA.8). Each prompt provides the task definition, inclusion and exclusion criteria, and output format. Together, they make the category definitions reproducible and provide the pivot from qualitative narrative taxonomy to synthetic calibration data.

English classification prompt (Demand Management / DM):

Task: Classify a fiscal measure based on its stated motivation. Does the measure correspond to Demand Management (DM)?

Definition of Demand Management (DM):

A measure belongs to this category when it is motivated by the need to change the short-run level or composition of aggregate demand in response to current output, inflation, unemployment, overheating, recession, or balance-of-payments conditions.

This includes:

- Tax increases intended to restrain consumption, imports, aggregate demand, or inflation.
- Tax cuts intended to support consumption, disposable income, investment demand, or economic activity during a slowdown.
- Measures addressing a balance-of-payments problem through demand restraint or import compression.
- Temporary or reversible measures justified by immediate cyclical conditions.

Do NOT classify as Demand Management if:

- The immediate objective is to close a current budget or external financing gap; classify it as Deficit Reduction.
- The measure gives short-run assistance to firms or a distressed sector through costs, liquidity, hiring, or production incentives; classify it as Supply Stimulus.
- The measure is a long-run structural reform intended to improve incentives, productivity, or competitiveness.
- The measure finances a specific spending program or implements a binding legal requirement.

Output format:

- Respond with one option:
- Yes - [short reasoning]
- No - [short reasoning]

Prompt OA.1: Classification Prompt for the Demand Management (DM) Category

English classification prompt (Supply Stimulus / SS):

Task: Classify a fiscal measure based on its stated motivation. Does the measure correspond to Supply Stimulus (SS)?

Definition of Supply Stimulus (SS):

A measure belongs to this category when it is motivated by short-run support for production, investment, employment, profitability, liquidity, or competitiveness, especially during a downturn or for a currently distressed sector.

This includes:

- Temporary tax relief intended to support firms, hiring, or investment during weak current conditions.
- Measures intended to protect the immediate liquidity or competitiveness of a sector.
- Tax cuts described as supporting recovery through short-run supply conditions.
- Temporary reductions in employment or business taxes intended to limit current job losses.

Do NOT classify as Supply Stimulus if:

- The justification is a durable structural or incentive reform whose purpose is long-run productivity or competitiveness; classify it as Long Run.
- The mechanism is short-run aggregate demand management rather than support for production or firms.
- The motivation is deficit reduction, spending finance, or legal compliance.
- The change is routine indexation or an administrative adjustment without a supply rationale.

Output format:

- Respond with one option:
- Yes - [short reasoning]
- No - [short reasoning]

Prompt OA.2: Classification Prompt for the Supply Stimulus (SS) Category

English classification prompt (Deficit Reduction / DR):

Task: Classify a fiscal measure based on its stated motivation. Does the measure correspond to Deficit Reduction or a current balance-of-payments crisis (DR)?

Definition of Deficit Reduction / current balance-of-payments crisis (DR):

A measure belongs to this category when it is motivated by a budget or external deficit caused by current conditions, an immediate need to raise or preserve revenue, a current reserve problem, or urgent government financing pressure.

This includes:

- Measures introduced to close a current budget deficit or meet an immediate Treasury financing need.
- Measures responding directly to a current external deficit or loss of reserves.
- Revenue measures responding to current investor-confidence or financial-market pressure on government financing.
- Temporary levies or restrictions of reliefs introduced to meet a documented near-term revenue shortfall.

Do NOT classify as Deficit Reduction if:

- The measure addresses a balance-of-payments problem by changing aggregate demand or imports; classify that mechanism as Demand Management.
- The measure implements an inherited or medium- or long-run fiscal consolidation strategy not forced by current conditions; classify it as Deficit Consolidation.
- The measure finances a specific new spending program; classify it as Spending-Driven.
- The stated reason is compliance with a pre-existing binding legal or institutional requirement; classify it as External.
- The document reports a revenue effect but does not state a current deficit, reserve, or financing rationale.

Output format:

- Respond with one option:
- Yes - [short reasoning]
- No - [short reasoning]

Prompt OA.3: Classification Prompt for the Deficit Reduction (DR) Category

English classification prompt (Spending-Driven / SD):

Task: Classify a fiscal measure based on its stated motivation. Does the measure correspond to Spending-Driven (SD)?

Definition of Spending-Driven (SD):

A measure belongs to this category when it is explicitly motivated by the need to finance, offset, or enable a specific increase in public spending or a named public program.

This includes:

- Tax increases introduced to cover the cost of a new or expanded program.
- Revenue measures described as necessary to finance a specific spending proposal.
- Exceptional levies earmarked for a stated public expenditure.
- Restrictions of tax reliefs intended to release resources for a targeted spending program.
- Charges introduced to recover the stated public cost of a specific service or program.

Do NOT classify as Spending-Driven if:

- The motivation is general current deficit reduction or medium-term consolidation.
- The measure is intended to manage aggregate demand.
- A spending program is mentioned but the tax change is not presented as financing or offsetting it.
- The stated reason is compliance with a binding legal or institutional obligation.

Output format:

- Respond with one option:
- Yes - [short reasoning]
- No - [short reasoning]

Prompt OA.4: Classification Prompt for the Spending-Driven (SD) Category

English classification prompt (Long Run / LR):

Task: Classify a fiscal measure based on its stated motivation. Does the measure correspond to Long Run (LR)?

Definition of Long Run (LR):

A measure belongs to this category when it is motivated by durable structural, supply-side, or incentive effects intended to improve productivity, investment, saving, competitiveness, or economic efficiency over time, rather than to offset a current shock.

This includes:

- Reforms intended to improve work, saving, investment, or entrepreneurial incentives over the long run.
- Measures intended to raise productivity, capital formation, or technological modernization.
- Durable reforms intended to improve efficiency or international competitiveness.
- Tax simplification and deregulation measures with a stated structural rationale.
- Long-term support for businesses or sectors that is not introduced in response to current distress.

Do NOT classify as Long Run if:

- The measure preserves the immediate liquidity, employment, or competitiveness of a distressed firm or sector; classify it as Supply Stimulus.
- The measure manages short-run demand, inflation, output, or employment.
- The measure responds to a current deficit or financing emergency.
- Competitiveness is mentioned only as a response to current external pressure.
- Simplification is purely administrative and no structural or incentive rationale is stated.

Output format:

- Respond with one option:
- Yes - [short reasoning]
- No - [short reasoning]

Prompt OA.5: Classification Prompt for the Long Run (LR) Category

English classification prompt (Ideological / IDEO):

Task: Classify a fiscal measure based on its stated motivation. Does the measure correspond to Ideological (IDEO)?

Definition of Ideological (IDEO):

A measure belongs to this category when it is motivated by a long-horizon social or political goal, a view about fairness or the proper design of the tax system, or a standing political commitment, independently of current macroeconomic performance and not to offset a current shock.

This includes:

- Measures justified by fairness, justice, social values, public health, environmental responsibility, or distributional principles.
- Reliefs or exemptions for groups described as deserving or vulnerable.
- Reforms implementing an electoral promise or standing political commitment.
- Anti-avoidance or anti-evasion measures when no current deficit, financing, or stabilization motive is stated.
- Changes intended to make the tax system fairer or more coherent as a matter of political or social philosophy.

Do NOT classify as Ideological if:

- The operative motivation is current demand management, deficit reduction, or financing pressure.
- The operative motivation is medium- or long-run fiscal consolidation.
- The measure is required by a pre-existing binding legal or institutional obligation.
- Anti-avoidance is used specifically to address a stated current revenue shortfall.
- The change is routine indexation or a neutral administrative adjustment without a stated social, political, or anti-avoidance objective.

Output format:

- Respond with one option:
- Yes - [short reasoning]
- No - [short reasoning]

Prompt OA.6: Classification Prompt for the Ideological (IDEO) Category

English classification prompt (External / EXT):

Task: Classify a fiscal measure based on its stated motivation. Does the measure correspond to External (EXT)?

Definition of External (EXT):

A measure belongs to this category when a pre-existing court ruling or binding legal or institutional obligation requires the government to change tax policy and leaves no contemporaneous discretion over whether to act. The operative motivation must be compliance with that requirement, not a response to current macroeconomic conditions.

This includes:

- Measures required to implement a court or judicial ruling.
- Measures required to enforce a binding EU or international directive.
- Measures required for compliance with a ratified treaty or binding international agreement when institutional compliance is the stated reason.
- Measures required by another pre-existing legal obligation that compels the tax change.

Do NOT classify as External if:

- The measure is merely enacted in a statute or finance bill; ordinary legislation does not by itself make a discretionary tax change External.
- The measure responds to a balance-of-payments problem, exchange-rate pressure, reserve loss, or an external financing shortage.
- The measure supports exports, international competitiveness, foreign trade, or investor confidence.
- The measure responds to a contemporaneous foreign or domestic macroeconomic shock.
- An international institution is mentioned but the operative motivation is demand management, current deficit reduction, or long-run reform.

Boundary examples:

- A tax change required to comply with a binding EU directive is External.
- A tax increase intended to reduce imports during a balance-of-payments crisis is Demand Management.
- A tax increase introduced to meet an immediate external financing shortfall is Deficit Reduction.
- A tax reform intended to improve international competitiveness over time is Long Run.

Output format:

- Respond with one option:
- Yes - [short reasoning]
- No - [short reasoning]

Prompt OA.7: Classification Prompt for the External (EXT) Category

English classification prompt (Deficit Consolidation / DC):

Task: Classify a fiscal measure based on its stated motivation. Does the measure correspond to Deficit Consolidation (DC)?

Definition of Deficit Consolidation (DC):

A measure belongs to this category when it is motivated by an inherited fiscal imbalance, a standing fiscal philosophy, or a medium- or long-run strategy to strengthen public finances, reduce borrowing, or lower public debt, rather than by a financing problem created by current conditions.

This includes:

- Measures implementing a multi-year fiscal consolidation plan.
- Structural revenue measures intended to improve the fiscal balance over several years.
- Measures addressing an inherited deficit or debt burden.
- Measures intended to offset the future fiscal cost of current policy actions.

Do NOT classify as Deficit Consolidation if:

- The measure responds to a current budget deficit, external deficit, reserve problem, or immediate financing pressure; classify it as Deficit Reduction.
- The revenue is introduced to finance a specific spending program; classify it as Spending-Driven.
- The measure is a short-run demand intervention or response to current output or inflation conditions.
- The stated reason is compliance with a pre-existing binding legal or institutional requirement.

Output format:

- Respond with one option:
- Yes - [short reasoning]
- No - [short reasoning]

Prompt OA.8: Classification Prompt for the Deficit Consolidation (DC) Category

OA.1.2.2 Synthetic Calibration Design

The calibration corpus is generated from the prompts defining the eight narrative categories above. For each category, I generate motivations under four complementary designs: canonical, institutional, boundary, and mixed-motive examples. The generated observations are self-contained, between 14 and 90 words, and exclude category names, classifier terminology, headings, and repeated templates. Boundary and mixed-motive examples retain potentially confounding or secondary considerations in the text, while a designated primary motivation determines the category label. The four generation designs are summarized below.

Table OA.4: Generation Designs for the Synthetic Calibration Samples

Design	Generation instruction
Canonical	Write clear positive examples for the target category in natural policy prose, without headings, category names, or repeated openings.
Institutional	Use the style of extracted U.S., Romer–Romer, or Cloyne policy motivations, including institutional references and policy language, while keeping every example synthetic.
Boundary	Include one misleading secondary cue from a nearby category while keeping the target category as the primary reason for the policy.
Mixed motive	Combine one primary target category with one secondary motivation in language resembling an actual policy document.

The following observations are taken verbatim from the stored calibration corpus and illustrate the differences across designs:

Canonical, long run. The Budget Statement framed the reduction in the corporate rate as a way to lift productivity by encouraging firms to renew plant, adopt newer methods, and commit capital with greater confidence over the long run.

Institutional, spending driven. The committee report states that the excise increase on tobacco products was adopted to provide dedicated receipts for the new federal health-insurance program beginning in fiscal 1974. The estimated yield was described as the financing source for premiums and administrative costs, not as a general budget device.

Boundary, demand management. The temporary cut in the sales levy was advanced to support household consumption while retail orders were weakening and factory shifts were being trimmed. Ministers noted that the measure could be withdrawn once inflation settled and private demand recovered. A separate note acknowledged the smaller receipts, but the immediate aim was to keep tills moving through the next few quarters.

Mixed motive, external. Following the Constitutional Court’s ruling that the inheritance tax unduly discriminated between direct descendants and more distant relatives, the government is obliged to amend the schedule of rates and exemptions within the timeframe set by the judgment. Ministers note that, although the reform will incidentally reduce planning opportunities and stabilize receipts, the core objective is to bring the law into conformity with constitutional equality guarantees and avoid further successful litigation.

For each design, I generate three independent samples using seeds 11, 29, and 47. Each sample contains 70 motivations per category, or 560 observations, yielding 6,720 synthetic motivations across the twelve calibration samples. The complete texts, category labels, secondary-category fields, generation design, and associated metadata are retained in the replication materials.

OA.1.2.3 Probability-Map Estimation

This subsection provides the estimation details behind the probability map summarized in Paper Section 3.1. Each of the twelve synthetic sets contains 70 motivations in each of the eight categories, for a balanced sample of 560 observations. Every synthetic motivation has one primary category, which provides its label. Examples may include secondary considerations to reproduce mixed policy language, but these are not treated as fractional labels.

For language model m and synthetic set k , let $h^{(m)}(M)$ denote the hidden-state representation of motivation M . I estimate a ridge-penalized multinomial logit mapping from this representation to the eight narrative categories:

$$\hat{P}_{m,k}(c \mid M) = \frac{\exp\left(\hat{\alpha}_{c,m,k} + h^{(m)}(M)' \hat{\lambda}_{c,m,k}\right)}{\sum_{c' \in \mathcal{G}} \exp\left(\hat{\alpha}_{c',m,k} + h^{(m)}(M)' \hat{\lambda}_{c',m,k}\right)}, \quad c \in \mathcal{G}.$$

The shrinkage parameter is selected by five-fold stratified cross-validation within each synthetic set. Applying the estimated map to historical motivation M_{ij} gives

$$p_{ij}^{(m,k)} = \sum_{c \in \mathcal{G}_{\text{exo}}} \hat{P}_{m,k}(c \mid M_{ij}),$$

where $\mathcal{G}_{\text{exo}}$ contains the categories associated with exogenous policy variation.

For each set-model pair, motivation-level probabilities are averaged within policy action:

$$p_i^{(m,k)} = \frac{1}{L_i} \sum_{j=1}^{L_i} p_{ij}^{(m,k)}.$$

The baseline policy-level probability averages over the twelve independently generated synthetic sets and the three language models:

$$p_i = \frac{1}{36} \sum_{m=1}^3 \sum_{k=1}^{12} p_i^{(m,k)}.$$

Set- and model-specific estimates are retained to assess how strongly the resulting support score depends on the synthetic sample and language-model representation. The associated diagnostics are reported below.

OA.1.2.4 Model Replication Details

The hidden-state representation stage and the open-weight motivation-extraction sensitivity exercise use the same pinned model versions. The exact language-model identity in these exercises is the pair consisting of the Hugging Face repository identifier and the resolved revision SHA: Runs used Python

Table OA.5: Language-model identifiers used for representation and extraction sensitivity

Short name	Hugging Face repository	Resolved revision
Gemma-4B	<code>google/gemma-4-E4B-it</code>	<code>d6436b3d62967e1af08bbb046c6300b2a9ae8e85</code>
Mistral-7B	<code>mistralai/Mistral-7B-Instruct-v0.2</code>	<code>63a8b081895390a26e140280378bc85ec8bce07a</code>
Qwen-9B	<code>Qwen/Qwen3.5-9B</code>	<code>c202236235762e1c871ad0ccb60c8ee5ba337b9a</code>

3.11.5, PyTorch 2.10.0+cu128, and a single NVIDIA L40S GPU.

OA.1.2.5 Alternative LLM Signal: Incremental Log Probabilities

The main text uses hidden-state calibration as the baseline way to extract category probabilities from language models. A simpler alternative is to use the same semantic prompts more directly. For motivation $M_{i\ell}$ and category $c \in \mathcal{G}$, let X_c be the category-specific prompt reported in Appendix OA.1.2.1.

Define the incremental log-probability signal

$$s_{i\ell,c}^{\log} := \log \Pr_{\text{LM}}(\text{Yes} \mid X_c \cup M_{i\ell}) - \log \Pr_{\text{LM}}(\text{Yes} \mid X_c). \quad (\text{OA.1})$$

The first term measures the model’s support for category c after observing the motivation; the second removes the prompt-specific baseline tendency to answer “Yes.” Collect these signals in

$$Z_{i\ell}^{\log} = \left(s_{i\ell,c}^{\log} \right)_{c \in \mathcal{G}}.$$

Because token-probability signals generated by different prompts need not be on a common probability scale, they should be treated as features to be calibrated rather than as final category probabilities. The same synthetic motivations used in the baseline calibration can provide the training sample: each synthetic motivation is generated from the category definitions in Appendix OA.1.2.1, so its category label is known by construction. One can therefore estimate, on this calibration sample, a multinomial mapping

$$\Pr(C_{i\ell} = c \mid Z_{i\ell}^{\log}) := \frac{\exp(\alpha_c + Z_{i\ell}^{\log \prime} \beta_c)}{\sum_{c' \in \mathcal{G}} \exp(\alpha_{c'} + Z_{i\ell}^{\log \prime} \beta_{c'})}, \quad c \in \mathcal{G}. \quad (\text{OA.2})$$

The resulting fitted probabilities can then be aggregated exactly as in Paper Section 2.1: sum the calibrated category probabilities over $\mathcal{G}_{\text{exo}}$ to obtain motivation-level support and average motivation-level support within the policy action to obtain p_i . This construction is useful as an alternative signal extraction method that keeps the prompt templates directly in the measurement equation. The main application instead uses hidden-state calibrated probabilities, reported in Paper Table 1, because the hidden-state route avoids relying on the model’s final Yes/No token probabilities as the primary signal.

OA.2 Supplementary Calibration Diagnostics

OA.2.1 Synthetic construction diagnostics

I audit the synthetic calibration corpus before using it to train the hidden-state classifier. These checks are construction diagnostics: they are designed to detect mechanical artifacts in the synthetic examples, not to validate the empirical identifying assumptions.

First, I verify exact balance. Each calibration set contains the same eight motivation categories and exactly 70 examples per category. This prevents category imbalance from mechanically affecting the multinomial mapping.

Second, I scan each motivation for prompt artifacts. The audit searches for banned prompt residues and template phrases such as “demand management motive”, “program funding motive”, “recovery through incentives”, “values-based reform”, “structural reform motive”, and related heading-like expressions. These checks are intentionally narrower than a general search for economically meaningful words: terms such as “long-term” or “deficit” may be legitimate content, while prompt headings or classifier language would indicate leakage from the generation prompt.

Third, I check for duplicate and near-duplicate examples. Exact duplicates are computed after lowercasing and normalizing punctuation and whitespace. Near duplicates are computed within each set and category using Jaccard similarity over normalized token sets, with the configured threshold equal to 0.72. This check is stricter than simply looking for identical strings and is meant to ensure that a category block is not populated by lightly reworded copies of the same example.

Fourth, I check for repeated openings. For each motivation, the audit records the first seven words and counts repeated openings within each set-category cell. The construction gate flags openings that occur more than three times within the same category and set. This prevents the classifier from exploiting formulaic first clauses rather than the substantive motivation.

Fifth, I estimate simple lexical shortcut diagnostics. For each synthetic set, I train cross-validated TF-IDF multinomial classifiers to predict the eight categories. One classifier uses the full motivation text; the other uses only the opening phrase. High full-text separability is expected in a synthetic category-labeled corpus and is reported only as a diagnostic. The more important shortcut check is the opening-only classifier: if the first few words alone predict category too well, the set may encode category through templates rather than motivation content.

Finally, I compare the synthetic texts with the extracted historical motivations. This is a coverage diagnostic, not a training step. The audit computes TF-IDF nearest-neighbor similarities between real motivations and synthetic examples, reports average and lower-tail similarities, and stores the least-covered real motivations for inspection. Real labeled observations are not added to the synthetic training set.

For the main robust synthetic corpus, all twelve generated sets are exactly balanced. The audit finds zero banned prompt-artifact rows, zero repeated-opening violations, zero exact duplicates, and zero near-duplicate pairs above the configured threshold. The opening-only classifier remains much less predictive than the full-text classifier, although a small number of generated sets are flagged as mild shortcut warnings. These warnings are retained as diagnostics and motivate the stability analysis across independently generated calibration sets.

OA.2.2 Dimensionality of the Hidden-State Representation

Each synthetic calibration set contains 560 observations, whereas the hidden-state representations contain 2,560 coordinates for Gemma-4B and 4,096 coordinates for Mistral-7B and Qwen-9B. The baseline multinomial logit addresses this dimensional imbalance through coefficient shrinkage selected by cross-validation. I additionally examine whether the resulting probabilities depend on the full dimension of the hidden-state representation.

For each calibration set and language model, I re-estimate the probability mapping after retaining either 32 or 64 principal components. Standardization and principal-component estimation are performed separately within each cross-validation training fold. Held-out observations are therefore

transformed using a principal-component basis estimated without them. After selecting the shrinkage parameter, I re-estimate the mapping on the complete synthetic set and apply it to the historical motivations. This procedure is repeated for all twelve calibration sets and three language models, and the resulting probabilities are aggregated exactly as in the baseline.

Table OA.6: Robustness to Hidden-State Dimensionality

<i>Panel A: Agreement with the full hidden-state probability</i>					
Sample	Specification	Pearson	Spearman	MAE	P90 abs. diff.
U.S. RR replication	PCA 64	0.995	0.991	0.022	0.044
	PCA 32	0.991	0.987	0.034	0.064
U.S. budget application	PCA 64	0.991	0.990	0.022	0.047
	PCA 32	0.986	0.985	0.028	0.061
Cloyne validation	PCA 64	0.990	0.984	0.017	0.039
	PCA 32	0.982	0.974	0.023	0.053
France VAT	PCA 64	0.989	0.985	0.019	0.041
	PCA 32	0.980	0.977	0.024	0.049
France non-VAT	PCA 64	0.988	0.985	0.019	0.040
	PCA 32	0.978	0.971	0.027	0.055

<i>Panel B: Separation in external narrative benchmarks</i>				
Specification	Romer–Romer		Cloyne	
	Exo–endo gap	AUC	Exo–endo gap	AUC
Full hidden state	0.176	0.699	0.101	0.655
PCA 64	0.166	0.712	0.096	0.659
PCA 32	0.152	0.700	0.089	0.652

Note: Panel A compares policy-level probabilities from each reduced-dimensional specification with the full-hidden-state baseline. Panel B reports the mean-support gap between externally classified exogenous and endogenous observations. Standardization and principal-component estimation are performed within each calibration fold.

Table OA.6 shows that dimensionality reduction preserves the ranking and levels of the historical support scores. The correlations with the baseline exceed 0.97 in every sample, and the external-benchmark gaps remain close to those obtained from the full hidden states. Figure OA.2 carries the same exercise through the support-indexed GMM estimator. The GDP and inflation responses are similar across the three representations.

This exercise isolates the role of representation dimensionality while holding fixed the synthetic motivations, category labels, language models, and aggregation rule. It does not test whether the synthetic categories fully span the historical language. Rather, it shows that the empirical results do not depend on retaining idiosyncratic directions in the full hidden-state representation.

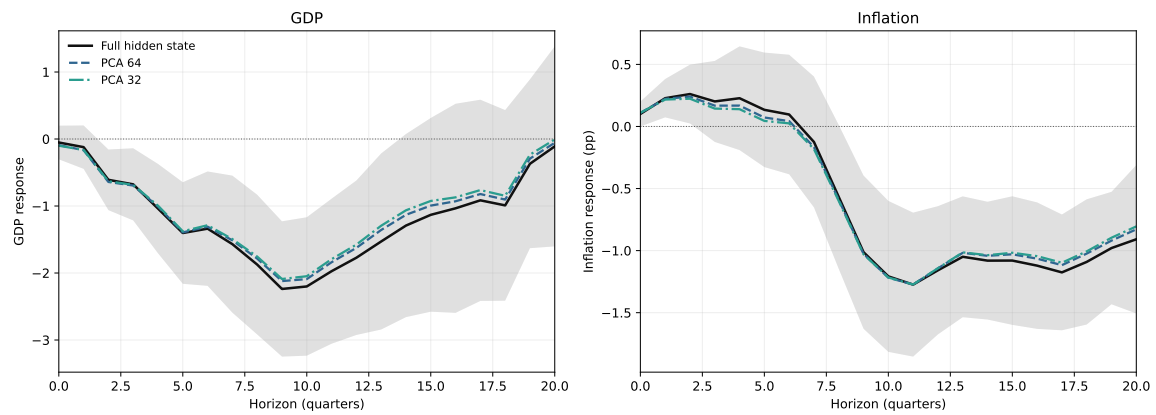


Figure OA.2: U.S. Budget Responses under Alternative Hidden-State Dimensions

Note: The black line reports the baseline affine $p = 1$ endpoint and the shaded area its 68% confidence interval. The dashed and dash-dotted lines re-estimate the same support-indexed GMM specification using probabilities constructed from 64 and 32 principal components, respectively.

OA.3 Supplementary Romer–Romer Evidence

OA.3.1 Probability Benchmark Settings

OA.3.1.1 Narrative extraction

This appendix documents the benchmark procedure behind Paper Section 5.1. The edited RR narratives are used to evaluate the support construction rather than for a direct binary relabeling exercise. Each example below shows the portion removed before the narrative is passed to the summarization stage. As in the main text, the deleted block includes the explicit RR classification sentence and the two immediately preceding sentences. Removing these passages prevents the model from seeing the authors’ bottom-line judgment while leaving enough narrative detail for the summarizer to recover multiple stated motivations from the remaining text.

Example 1: an endogenous measure

The motivation for the Revenue Act of 1951 was again the increase in spending related to the Korean War. The *1951 Economic Report* stated: “These new taxes are required to finance the defense effort; and to help keep total spending within the capacity of current production, so that inflation does not reduce the purchasing power” (p. 17). This same sentiment was echoed in a number of presidential speeches. On February 2, 1951, Truman said, “If we do not tax ourselves enough to pay for defense expenditures, the Government will ... add to total purchasing power and inflationary pressures” (*Special Message to the Congress Recommending a “Pay as We Go” Tax Program*, p. 2). These quotations make it clear that the tax increase was designed to keep growth normal.

It stated: “The military action in Korea, coupled with the general threat to world peace, has made it necessary to provide extraordinary increases in revenues to meet essential national defense expenditures” (82d Congress, 1st Session, House of Representatives Report No. 586, 6/18/51, p. 1). It also stressed the temporary nature of the tax, saying: “The 12½-percent flat across-the-board increase [in individual income tax rates] was selected as the form of increase, since it is not intended that this increase will be permanent, and therefore, it was desired to provide an increase which would not become an integral part of the rate structure” (p. 8). Because the tax increase was designed to counteract the effects of higher military spending and thereby keep growth normal, I classify it as an endogenous, spending-driven action.

Truman’s speech upon signing the bill stated that it would raise about \$5.5 billion of additional revenues at an annual rate (*Statement by the President Upon Signing the Revenue Act of 1951*, 10/20/51, p. 1). Both the *1952 Economic Report* (p. 135) and the *1951 Treasury Annual Report* (p. 44) gave the figure as \$5.4 billion. I follow my usual practice and use the number from the *Economic Report*.

The increase in corporate taxes was retroactive to April 1, 1951, while the remainder took effect on November 1, 1951 (*1951 Treasury Annual Report*, p. 51). Because the action took place before the middle of the quarter, I date the tax increase in 1951Q4. Approximately \$2.3 billion of the tax increase was from corporate taxes (p. 501). Therefore, the retroactive portion of the bill increased taxes at an annual rate of \$2.3 billion · 2 extra quarters, or \$4.6 billion. Adding this to the steady-state effect of \$5.4 billion yields a tax increase of \$10 billion in 1951Q4. The return to the steady-state level of tax increase implies a tax cut of \$4.6 billion in 1952Q1. If one ignored the retroactive part of the legislation, there would be just a tax increase of \$5.4 billion in 1951Q4.

The tax increase largely took the form of an increase in marginal rates. The act also raised the capital gains tax, the tax on corporate profits, and some excise taxes (*1951 Treasury Annual Report*, pp. 50–52). Most of the changes were explicitly temporary. The individual income tax increases were legislated to expire on January 1, 1954; the corporate and excise tax increases were to continue until March 31, 1954 (*1951 Treasury Annual Report*, pp. 51–52).

Example 2: an exogenous measure

This bill was a significant change of the income tax system. It was the first comprehensive revision of the Internal Revenue Code in 75 years (*Statement by the President Upon Signing Bill Revising the Internal Revenue Code*, 8/16/54, p. 1).

This bill was passed primarily to increase the fairness and efficiency of the tax system. It had been under discussion since 1952 (Holmans, 1961, p. 219), well before the 1953–54 recession. At the signing ceremony, Eisenhower stressed: “It is a law which will help millions of Americans by giving them fairer tax treatment than they now receive.” He also said, “economic growth will be fostered by such provisions as more flexible depreciation and better tax treatment of research and development costs, thus encouraging all business—large and small—to modernize and expand” (*Statement by the President Upon Signing Bill Revising the Internal Revenue Code*, 8/16/54, p. 1). The Report of the Committee on Ways and Means to Accompany H.R. 8300 (Internal Revenue Code of 1954) repeated this motivation. It stated: “In general, the purpose of these changes has been to remove inequities, to end harassment of the taxpayer and to reduce tax barriers to future expansion of production and employment” (83d Congress, 2d Session, House of Representatives Report No. 1337, 3/9/54, p. 1).

From this perspective, the reduction in revenues at a given level of income was the result of a belief that the bill would stimulate growth and of a desire to make the reform roughly revenue neutral after accounting for the growth effects. The *1954 Economic Report* stated, “A number of additional tax measures should now be taken in order to strengthen the forces of growth in employment and production. These measures, which involve some immediate sacrifices of revenue, contain the seeds of important future revenue gains to be reaped from the economic growth they will stimulate” (p. 79). The *1955 Economic Report* described the motivations for the bill as “to stimulate competitive enterprise, to strengthen the floor of security for the individual, and to curb tendencies toward either depression or inflation,” and referred to “the loss of 1.4 billion dollars as a result of enacting structural tax changes” as almost a side effect of the reform (pp. 19–20).

The possible countercyclical benefits of the bill were certainly mentioned frequently. As discussed above, the administration favored some additional tax relief in mid-1954 because of further declines in defense spending. And, much of the Congressional debate focused on the need for additional stimulus. Indeed, several amendments were offered that involved increasing the personal exemption and cutting tax rates (Holmans, 1961, pp. 229–233). However, once the Excise Tax Reduction Act of 1954 was passed, the Administration argued strenuously against further tax cuts, especially the increase in the personal exemption.

[1] On March 15, 1954, Eisenhower said: “at this time economic conditions do not call for an emergency program that would justify larger Federal deficits and further inflation through large additional tax reductions” (*Radio and Television Address to the American People on the Tax Program*, 3/15/54, p. 4; see also *Annual Budget Message to the Congress: Fiscal Year 1955*, 1/21/54, p. 8). Also, the House report on the bill explicitly downplayed the countercyclical aspects, saying: “This bill is a long overdue reform measure which is vitally necessary regardless of the momentary economic conditions and should not be confused with other measures which may be, or might become, appropriate in the light of a particular short-run situation” (House Report No. 1337, pp. 1–2). Thus, it seems clear that at least some of the tax cuts contained in the Excise Tax Reduction Act of 1954 and the Internal Revenue Code of 1954 should be classified as exogenous.

[2] Thus, it seems clear that at least some of the tax cuts contained in the Excise Tax Reduction Act of 1954 and the Internal Revenue Code of 1954 should be classified as exogenous. There was a cut beyond what the administration thought was justified by the decline in spending and the state of the economy. Because the excise tax reduction came first and had no obvious other motivation, I choose to classify it as endogenous and the Internal Revenue Code of 1954 as an exogenous, long-run action.

[3] There was a cut beyond what the administration thought was justified by the decline in spending and the state of the economy. Because the excise tax reduction came first and had no obvious other motivation, we choose to classify it as endogenous and the Internal Revenue Code of 1954 as an exogenous, long-run action. We feel this makes sense given that the Internal Revenue Code of 1954 clearly had a substantial exogenous component and indeed was passed after it was widely understood that the trough of the recession (May 30, 1954) had been turned (Holmans, 1961, p. 233).

[4] Because the excise tax reduction came first and had no obvious other motivation, we choose to classify it as endogenous and the Internal Revenue Code of 1954 as an exogenous, long-run action. We feel this makes sense given that the Internal Revenue Code of 1954 clearly had a substantial exogenous

component and indeed was passed after it was widely understood that the trough of the recession (May 30, 1954) had been turned (Holmans, 1961, p. 233). But, at some level, which of the two tax cuts is classified as endogenous and which as exogenous is somewhat arbitrary.

The act was signed in mid-August 1954, so I date the initial tax change in 1954Q3. My sources give a figure of $-\$1.4$ billion for the revenue effects at an annual rate (for example, *1955 Economic Report*, p. 20; *1954 Treasury Annual Report*, p. 44). The bill was retroactive to January 1, 1954 (*1954 Treasury Annual Report*, p. 286). It thus reduced taxes by $3 \cdot \$1.4$ billion, or $\$4.2$ billion at an annual rate, in 1954Q3. The return to the steady-state reduction of $\$1.4$ billion in 1954Q4 implies a tax increase of $\$2.8$ billion in 1954Q4. If one chose to neglect the retroactive element, there would be just a tax cut of $\$1.4$ billion in 1954Q3.

The changes consisted largely of the removal of obvious inequities, complications, ambiguities, and opportunities for tax avoidance, and of provisions designed to stimulate investment. All the changes were legislated to be permanent.

OA.3.1.2 Overlap Between the RR Series and Budget-Based Measures

The Romer–Romer series and the budget-based data cover different parts of the federal tax-policy process and are therefore not expected to coincide observation by observation. Romer and Romer record enacted federal tax changes, whereas the budget data record tax proposals made by the administration to Congress in the annual *Budget of the United States Government*. Some overlap nevertheless arises. A proposal reported in the Budget may subsequently be enacted and enter the Romer–Romer series, and later budgets may propose extensions, accelerations, reductions, or other modifications of provisions associated with earlier Romer–Romer episodes.

Relating the two sources therefore requires more than matching law titles. A single Romer–Romer episode may contain several tax provisions that appear separately in the budget data, while a budget proposal may concern only one component of a broader piece of legislation. In other cases, a statute appears in a later Budget because the administration proposes to modify one of its provisions rather than because the Budget records the original tax change. I therefore compare the two sources at the level of individual tax provisions, taking account of the tax instrument, proposed effective date, and reported revenue effect.

A preliminary comparison of the 50 Romer–Romer episodes with the budget-based measures identifies a small number of relatively clear overlaps. Two examples are reported in Table OA.7.

Table OA.7: Examples of RR–Budget Overlap

RR episode	Budget-based measure	Comparison
Revenue and Expenditure Control Act of 1968	The budget data include the temporary 10 percent surcharge on individual and corporate income taxes, together with related excise-tax extensions.	The same broad tax action appears in both sources, although the budget data record the individual provisions separately.
Omnibus Budget Reconciliation Act of 1993	The fiscal-year 1994 Budget includes the proposed high-income individual income-tax increase associated with the legislation.	The individual-income-tax component can be matched relatively cleanly. Other components of OBRA 1993 are less straightforward to reconcile.

Several other episodes have plausible but less direct counterparts. For example, provisions associated with the Economic Growth and Tax Relief Reconciliation Act of 2001 appear in later budget proposals through accelerations of previously legislated tax changes. These observations are related to the same legislation but do not correspond to the original Romer–Romer timing. The Economic Recovery Tax Act of 1981, the Tax Equity and Fiscal Responsibility Act of 1982, the Revenue Act of 1978, and OBRA 1990 also have related entries in the budget documents, although the correspondence

is often limited to particular provisions or revenue components. For the Revenue Acts of 1950 and 1951, the candidate budget entries are not sufficiently close to the enacted provisions to treat them as direct matches.

For this reason, I treat the budget-based data as a separate source of tax-policy variation rather than as observations that can simply be appended to the Romer–Romer series. Constructing a combined series would require a provision-level crosswalk linking each Romer–Romer action to the corresponding budget proposal, where such a counterpart exists.

OA.3.2 Estimator-Specific Romer–Romer Distributed-Lag Diagnostic

As a complementary benchmark, Figure OA.3 reports the exercise closest to the published Romer–Romer dynamic specification. Quarterly real GDP growth is regressed on the contemporaneous tax shock and twelve lags of that shock, together with eleven lags of GDP growth, over the 1950Q1–2007Q4 sample. The figure reports the simulated cumulative GDP response implied by this distributed-lag equation after a one-percent-of-GDP tax innovation. The inflation panel uses the same original-versus-high-support comparison in the baseline inflation projection. This check preserves the original RR distributed-lag specification and therefore answers a different question from the harmonized binary-selection-versus-probabilistic moment comparison in Paper Figure 3.

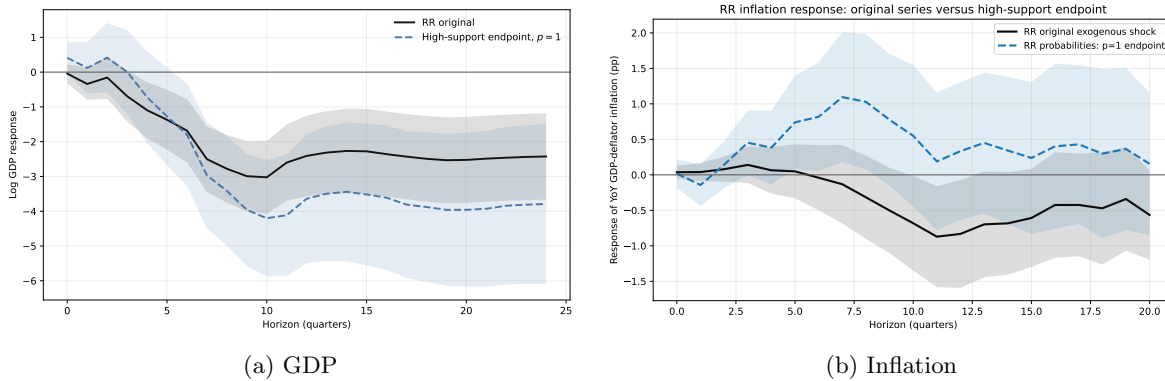


Figure OA.3: Romer–Romer Responses: Original Series versus High-Support Endpoint

Note: Panel (a) reports the exact Romer–Romer GDP specification. Panel (b) reports the corresponding inflation projection using year-over-year GDP-deflator inflation. The black line is the original RR exogenous shock series. The colored line is the affine two-type estimator evaluated at $p = 1$. Shaded areas denote 68% confidence bands.

The original RR series reaches a peak contraction of roughly 3 log points around quarter 10, in line with the original Romer–Romer result. The two-type $p = 1$ endpoint uses the probability ranking to estimate the high-support endpoint of an affine mixture, rather than converting the continuous score back into a binary rule. Its proximity to the original RR response under this distributed-lag estimator should not be read as superseding the harmonized moment comparison, which uses a different macroeconomic estimator. Figure OA.4 reports the corresponding low- and high-support endpoint diagnostic.

OA.3.3 Two-type endpoint diagnostic

Figure OA.4 unpacks the two-type estimator used in the exact-RR distributed-lag exercise. The left panel reports the two endpoint responses, $p = 0$ and $p = 1$. The right panel shows the implied linear mapping at horizon $h = 10$, where the exact-RR benchmark reaches its peak contraction in the baseline specification. Figure OA.5 reports the corresponding diagnostic for inflation using the baseline inflation projection.

OA.3.4 Narrative Ambiguity in the Romer–Romer Probability Benchmark

This appendix provides a qualitative audit of the policy-level probabilities used in the Romer–Romer benchmark. The exercise uses the edited narrative texts described in Appendix OA.3.1.1: the model observes the policy narrative after the explicit Romer–Romer classification statement and the immediately adjacent classification cues have been removed. The purpose is not to mechanically reproduce the original binary label in Romer and Romer (2010). Instead, the diagnostic asks whether intermediate

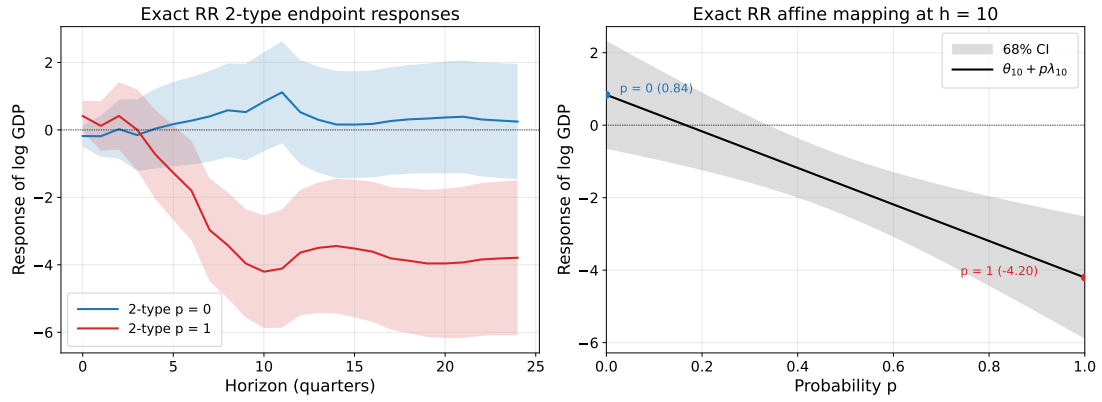


Figure OA.4: Two-Type RR Benchmark: Endpoint IRFs and the Mapping at $h = 10$

Note: The left panel reports the exact-RR two-type endpoint responses for $p = 0$ and $p = 1$, each with 68% confidence bands. The right panel plots the fitted affine response as a function of p at horizon $h = 10$, together with its 68% confidence band.

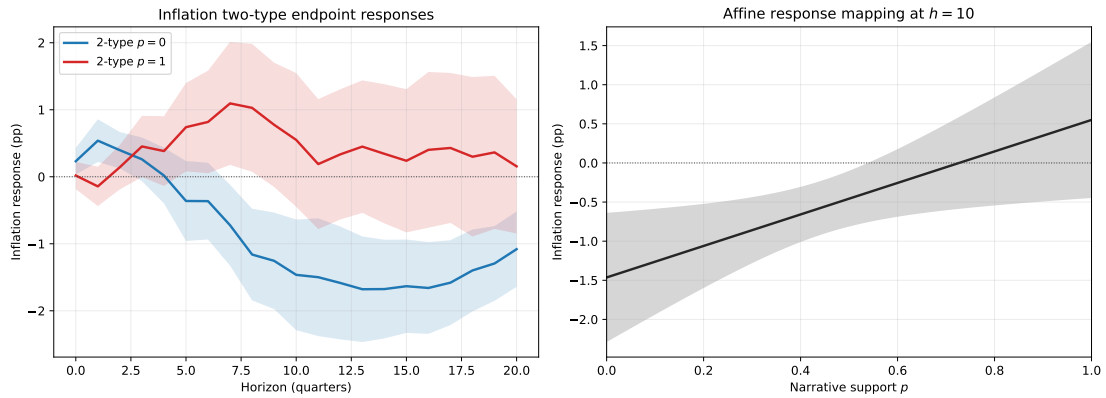


Figure OA.5: Inflation Two-Type RR Benchmark: Endpoint IRFs and the Mapping at $h = 10$

Note: The left panel reports the inflation two-type endpoint responses for $p = 0$ and $p = 1$, each with 68% confidence bands. The right panel plots the fitted affine response as a function of p at horizon $h = 10$, together with its 68% confidence band. The panels report the inflation counterpart of the GDP endpoint analysis.

values of $\hat{p}_i$ correspond to genuine ambiguity in the narrative record, or whether they reveal cases where the estimated support appears too high or too low relative to the semantic object used in the narrative-tax literature.

The key distinction is that the Romer–Romer label combines several judgment margins. Some episodes are excluded because the tax change responds to current or prospective macroeconomic conditions; others because it is tied to spending or benefit changes. Conversely, some deficit-driven actions are retained as exogenous when they are not judged to respond to current cyclical conditions. The probability measure applies the paper’s common taxonomy instead: current-macro responses, spending-package dependence, and fiscal financing motives can contribute different amounts of policy-response support. Differences from the RR label are therefore informative about category boundaries rather than, by themselves, evidence of measurement failure. Table OA.8 summarizes the main cases.

Table OA.8: Romer–Romer Policy Narratives and Estimated Support

Policy	RR category	$\hat{p}_i$	Interpretation
Expiration of Excess Profits Tax and of Temporary Income Tax Increases	Endogenous, spending-driven	0.306	The low support is consistent with the original narrative: the tax reductions followed the end of Korean-War spending and a changed postwar macro-fiscal environment.
Social Security Amendments of 1958	Endogenous, spending-driven	0.651	The RR exclusion is about benefit financing and actuarial balance, not a direct response to output or inflation. The intermediate support is sensible if p_i primarily captures macro exogeneity rather than all package contamination.
Revenue Act of 1964	Exogenous, long-run	0.448	A boundary case. RR emphasize long-run growth and explicitly reject a quick recession-prevention tax cut. The same text, however, invokes unemployment, below-capacity production, and fiscal drag.
Revenue and Expenditure Control Act of 1968	Endogenous, countercyclical	0.208	The narrative is directly macro-stabilization oriented: overheating, inflation pressure, fiscal restraint, and external pressure on the dollar.
Tax Reform Act of 1969	Endogenous, countercyclical	0.640	The probability is high relative to a strict RR rule. The narrative contains inflation-control language, but also fairness, equity, poverty relief, and tax-reform motives.
Crude Oil Windfall Profit Tax Act of 1980	Exogenous, long-run	0.464	RR treat the tax as unrelated to macro stabilization. The intermediate support reflects oil-price decontrol, energy policy, and consumer-assistance uses of revenue.
Social Security Amendments of 1983	Exogenous, deficit-driven	0.690	Support remains below one because the narrative is strongly tied to trust-fund solvency, Social Security financing, and deficit reduction.
Tax Reform Act of 1986	Exogenous, long-run	0.950	A clean alignment case: tax reform, fairness, simplicity, and long-run incentives dominate the narrative, with little role for current macro conditions.
Omnibus Budget Reconciliation Act of 1990	Exogenous, deficit-driven	0.619	RR classify the act as deficit-driven, but the narrative acknowledges recession timing and weak expected growth.
Omnibus Budget Reconciliation Act of 1993	Exogenous, deficit-driven	0.657	The act is not a short-run stabilization measure, but deficit reduction is the central state-contingent motive.
Taxpayer Relief Act of 1997 and Balanced Budget Act of 1997	Endogenous, spending-driven	0.713	The RR exclusion reflects budget-package contamination. The tax-side rationale is less clearly cyclical, so estimated support is relatively high.
Economic Growth and Tax Relief Reconciliation Act of 2001	Endogenous, countercyclical	0.515	The intermediate support reflects mixed motives: surplus return and long-run incentives on one side, explicit slowdown and short-run stimulus language on the other.

Note: RR category denotes the original Romer–Romer classification. Probabilities use the baseline taxonomy and are averaged across calibration samples and representation models.

Clean alignment. When the narrative evidence is relatively clear, the probability moves in the expected direction. The Tax Reform Act of 1986 receives very high support, $\hat{p}_i = 0.950$. This is consistent with the Romer–Romer narrative, where the motivation is tax reform, fairness, simplicity, and long-run incentives. The narrative does mention economic growth, but the relevant growth motive

is long-run reform rather than contemporaneous stabilization.

At the other end, the expiration of the Korean-War temporary tax increases receives very low support, $\hat{p}_i = 0.306$. Romer–Romer treat the expiration as a deliberate policy choice after the wartime fiscal environment changed. The motivation is tied to the end of Korean-War spending and to the macro-fiscal context in which temporary taxes were allowed to lapse. The low probability is therefore not a failure to recover a label; it reflects the same identifying concern that motivates the original endogenous, spending-driven classification.

The 2001 Economic Growth and Tax Relief Reconciliation Act receives intermediate support, $\hat{p}_i = 0.515$. The act combines an argument about returning projected budget surpluses and improving long-run incentives with an explicit short-run stimulus motive during a slowdown. The probability records both exogeneity-supporting and policy-response elements.

Boundary exogenous cases. The Revenue Act of 1964 is the central boundary case. Romer–Romer classify it as exogenous and long-run because the stated objective was to raise the economy’s longer-run growth path rather than to prevent an immediate recession. In their narrative, the administration explicitly distinguished the proposal from a “quick, temporary tax cut” and emphasized longer-run economic challenges. At the same time, the surrounding text contains substantial language about unemployment, below-capacity production, and the removal of fiscal drag. The resulting probability, $\hat{p}_i = 0.448$, records partial rather than complete documentary support for the exogenous interpretation.

This case is also empirically important because it helps explain the non-monotone threshold diagnostic. With support below 0.5, the 1964 tax-cut episode is excluded at each reported cutoff, even though it is central to the original Romer–Romer contractionary response. Other medium-support episodes enter or leave as the cutoff changes. A higher cutoff therefore does not simply remove uniformly less credible observations; it changes the composition of the historical episodes identifying the response.

The Crude Oil Windfall Profit Tax Act of 1980 illustrates a different kind of boundary case. Romer–Romer classify the act as exogenous because the tax was motivated by capturing windfall profits from oil-price decontrol rather than by managing output or inflation. The probability is nevertheless only 0.464. The policy is tied to oil-price regulation, energy policy, consumer assistance, and the use of earmarked revenues. Those features are not the same as current-macro endogeneity, but they make the tax change less like an isolated exogenous tax change.

Deficit and Social Security cases. Deficit-driven and Social Security cases reveal the main boundary between the two classification objects. Romer–Romer retain some deficit-driven tax increases as exogenous when the stated motivation is not to offset current cyclical conditions. Under the paper’s taxonomy, an action can nevertheless receive policy-response support when it is explicitly motivated by trust-fund financing, a fiscal imbalance, or a spending package. Intermediate probabilities in these cases reflect this broader set of possible links to the policy rule.

The Social Security Amendments of 1983 make this tension clear. Romer–Romer classify the episode as exogenous, deficit-driven, but the narrative is organized around Social Security solvency, trust-fund financing, and deficit reduction. The estimated support is $\hat{p}_i = 0.690$. This is above one half but remains well below full support despite the original exogenous classification. The tax change is presented as a response to a program-financing problem rather than as an isolated tax-policy experiment. The RR label is natural under a non-countercyclical criterion; the paper’s broader taxonomy assigns partial policy-response support to the financing motive.

The 1990 and 1993 budget acts show the same issue. OBRA 1990 receives $\hat{p}_i = 0.619$. The original classification is deficit-driven, but the narrative also notes recession timing and weak expected growth. OBRA 1993 receives $\hat{p}_i = 0.657$. The motive is deficit reduction to improve long-run macroeconomic performance, not short-run demand management. Still, deficit reduction is a state-contingent fiscal motive. The medium probabilities therefore reflect the fact that RR’s deficit-driven convention is not the same object as full narrative support for an orthogonal tax shock.

The Social Security Amendments of 1958 point in the opposite direction. They are classified as endogenous, spending-driven, because higher payroll taxes financed higher benefits and improved actuarial balance. The estimated support is 0.651, which is high relative to the hard RR label. The narrative concerns benefit financing rather than output, inflation, or stabilization. The probability may therefore be too high for the full RR exclusion rule but remains consistent with support for macro exogeneity.

Departures that reveal category boundaries. The Tax Reform Act of 1969 is a case where the probability may be high relative to a strict reading of the RR taxonomy. Romer–Romer classify the act as endogenous, countercyclical, partly because the administration accepted a period of slower growth to reduce inflation. The extracted motivations also include fairness, poverty relief, and political tax relief. The probability, $\hat{p}_i = 0.640$, places substantial weight on the non-cyclical components. Under a strict exclusion of policies motivated by current or prospective macro conditions, the inflation-control motive would instead call for lower support.

The 1997 Taxpayer Relief Act and Balanced Budget Act provide a complementary example. Romer–Romer classify the tax cuts as endogenous, spending-driven, because they are tied to a broader budget package in which spending reductions made the tax cuts feasible. The estimated support, $\hat{p}_i = 0.713$, is high relative to that hard label. The reason is that the tax-side motive is not clearly a short-run stabilization response. Under the RR exclusion rule, however, the package structure is itself a source of contamination. A probability focused on macro exogeneity may therefore under-penalize spending-package dependence unless that distinction is made explicit in the measurement target.

Takeaway. Estimated support is not a noisy approximation to the Romer–Romer hard label. High-support cases are dominated by tax reform and long-run motives, whereas low-support cases feature war spending, macroeconomic stabilization, or explicit short-run stimulus. Departures arise mainly where the RR convention and the paper’s taxonomy treat deficit financing, social-insurance financing, or spending-package dependence differently. The probability measures documentary support for exogeneity under the common taxonomy; it is not an automated replication of the original classification.

OA.3.5 Training-data leakage robustness check

To assess whether the language model has memorized specific source material, I implement a sentence-reconstruction test taken from Golchin and Surdeanu’s (2024) contamination detection framework. The intuition is the following: if a model has encountered a document during training, it may be able to reproduce missing content from that document with unusual precision—especially when provided with cues pointing to the original source. To implement this test, I systematically remove sentences from the Romer and Romer narratives and asks the model to reconstruct them under two conditions: one relying solely on the surrounding context, and the other explicitly guiding the model by referencing the source paper and section. If the model performs significantly better under guided prompting, this may suggest prior exposure to the original text. To quantify and evaluate the reconstruction quality, I use both semantic (embedding cosine similarity) and lexical (ROUGE-L) metrics, followed by a bootstrap-based significance test to flag cases of potential memorization.

Using both metrics, and evaluating with `gpt-4o-mini` and `gemini-2.0-flash`, the test flags no narrative as a potential case of memorization. This finding provides no evidence of detectable sentence-level reconstruction under the tested models, prompts, and metrics. It does not establish that the source paper was absent from the models’ training data or rule out other forms of training-data exposure.

Procedure.

1. **Multiple draws & prompts.** For each narrative d I perform $T = 10$ draws. In draw t , one sentence $s_{d,t}$ is chosen uniformly at random, removed, and replaced by the token [MISSING

SENTENCE]. I then issue two prompts a general prompt, and a guided prompt (Prompts OA.9).

General prompt:

You are provided with a passage in which one sentence has been removed and replaced with [MISSING SENTENCE]. Based ONLY on the context provided, generate a plausible sentence that could fill the gap. You do not need to match any specific source text — just ensure the result is coherent and contextually appropriate.

{narrative text from Romer and Romer}

Reconstructed sentence:

Guided prompt:

You are provided with a passage from the scientific paper *A Narrative Analysis of Postwar Tax Changes* by Christina D. Romer and David H. Romer, published in 2010. The passage comes from the section titled ‘{title}’ in the main body of the paper.

One sentence has been removed and replaced with [MISSING SENTENCE]. Based **only** on the context provided and your knowledge of this specific paper, reconstruct the missing sentence exactly as it appeared.

{narrative text from Romer and Romer}

Reconstructed sentence:

Prompt OA.9: General and Guided Prompt

The model returns two reconstructions $\hat{s}_{d,t}^{\text{gen}}$ and $\hat{s}_{d,t}^{\text{guide}}$.

2. **Similarity metrics.** For each original $s_{d,t}$ and each reconstruction, compute:

- *Embedding cosine similarity* $\rho_{d,t}^{\text{gen}} = \cos(s_{d,t}, \hat{s}_{d,t}^{\text{gen}})$, $\rho_{d,t}^{\text{guide}} = \cos(s_{d,t}, \hat{s}_{d,t}^{\text{guide}})$, using `text-embedding-3-large`.
- *ROUGE-L F1 score* $r_{d,t}^{\text{gen}}$ and $r_{d,t}^{\text{guide}}$ between the original and each reconstruction.

3. **Paired bootstrap test.** For each narrative d , define the mean differences

$$\Delta\rho_d = \frac{1}{T} \sum_{t=1}^T (\rho_{d,t}^{\text{guide}} - \rho_{d,t}^{\text{gen}}), \quad \Delta r_d = \frac{1}{T} \sum_{t=1}^T (r_{d,t}^{\text{guide}} - r_{d,t}^{\text{gen}}).$$

We perform a nonparametric paired bootstrap with $B = 1,000$ resamples of the T pairs $\{(\rho_{d,t}^{\text{guide}}, \rho_{d,t}^{\text{gen}})\}$ (and similarly for $\{(r_{d,t}^{\text{guide}}, r_{d,t}^{\text{gen}})\}$), yielding bootstrap distributions of $\Delta\rho_d$ and Δr_d . Each indicator $\mathbf{1}\{\Delta\rho_d^{*(b)} \leq 0\}$ (or $\mathbf{1}\{\Delta r_d^{*(b)} \leq 0\}$) counts a bootstrap sample in which the similarity under the guided prompt is no better than—or worse than—that under the general prompt; that is, the model failed to improve reconstruction quality despite being explicitly cued about the source. The empirical p-values are

$$p_d^\rho = \frac{1}{B} \sum_{b=1}^B \mathbf{1}\{\Delta\rho_d^{*(b)} \leq 0\}, \quad p_d^r = \frac{1}{B} \sum_{b=1}^B \mathbf{1}\{\Delta r_d^{*(b)} \leq 0\}.$$

Decision rule. A narrative d is flagged as potentially memorized if its bootstrap p -value for either similarity metric falls below 0.05, i.e. if guided prompting produces significantly closer reconstructions than general prompting.

OA.3.6 Differences Relative to the Romer–Romer Classification

The cases below describe differences between the constructed support measure and the Romer–Romer classification. They need not be measurement errors: several reflect genuine ambiguity in the documentary record or category boundaries that require the bespoke judgment used in the original narrative study. A common pattern is that evidence supporting an endogenous policy response is easier to state directly than evidence that a policy action is unrelated to current or anticipated macroeconomic conditions. In such cases, the support measure can place substantial probability on both interpretations rather than reproduce a binary label.

Steering growth above normal levels. A recurring difference concerns tax cuts intended to raise growth above normal levels. In at least three cases, the presence of growth-oriented language creates support for a countercyclical interpretation. Romer and Romer’s decision rule instead treats aspirations to lift growth above normal levels as exogenous when they are not tied to a negative output gap. These cases illustrate a category boundary that can be made more explicit in the prompt design or calibration examples.

Conflicting motivations. Other cases contain mixed motivations. For example, one narrative discusses a tax increase intended primarily for deficit reduction and solvency of the Highway Trust Fund, while also referring to inflation concerns and contemporaneous spending increases. The original classification treats the measure as exogenous but notes the ambiguity. The constructed support measure retains both elements of the documentary record rather than resolving the conflict through a single binary decision.

“Package” related measure. Finally, some cases hinge on bespoke decision rules. A salient example is the Tax Adjustment Act of 1966. Although it was introduced to counteract high growth following an exogenous tax cut, Romer and Romer classify it as exogenous, arguing that the alternative classification would yield nonsensical results. This context-specific, precedent-based reasoning is not fully summarized by the general category definitions.

OA.4 Additional U.S. Composition and Transmission Evidence

OA.4.1 Affine GMM Endpoint Diagnostic

The main text reports the fitted response at $p = 1$ under the baseline affine GMM estimator. Figure OA.6 reports the corresponding fitted value at $p = 0$ as a diagnostic. This contrast is useful for describing the affine specification: the contractionary medium-run response is concentrated at $p = 1$, while the fitted response at $p = 0$ is much flatter and eventually positive. The latter should not be read as a separately observed class of purely endogenous policies. Under the baseline convention, it is anchored jointly by no-action quarters assigned $p_t = 0$ and by recorded policy quarters receiving little support for exogeneity. Paper Appendix F shows how the fitted response at $p = 1$ changes when no-action quarters are placed elsewhere on the support scale or excluded from the support-indexed projections.

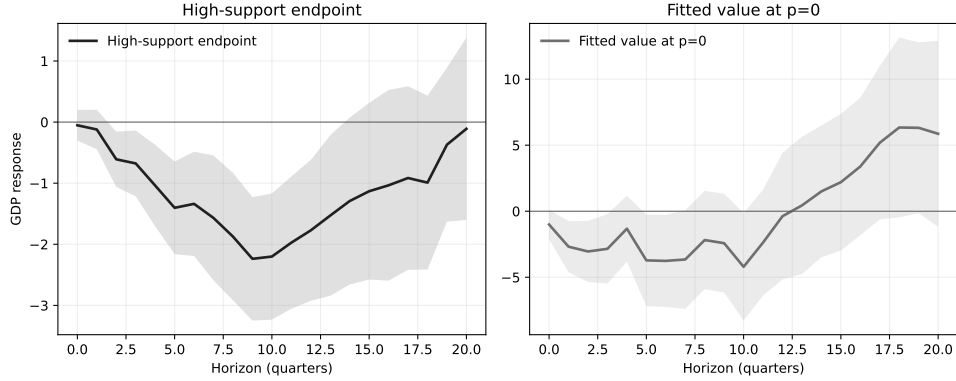


Figure OA.6: Affine GMM Endpoint Diagnostic: $p = 0$ and $p = 1$

Note: The figure reports the U.S. budget affine GMM fitted values for real GDP at $p = 1$ and $p = 0$. Both use the same baseline estimator, residualization, and fitted squared-tax denominator evaluated at the corresponding value of p , with pointwise 68% fully stacked HAC confidence bands. Under the baseline convention, the $p = 0$ fitted value is anchored by both no-action quarters and recorded policy quarters with little support for exogeneity; it is not a separately observed policy class.

OA.4.2 Component Diagnostics and Romer–Romer Comparison

Figure OA.7 reports the remaining component panels corresponding to Figure 14 in Romer and Romer (2010). The comparison is useful for separating the mechanism that carries over from the original RR exercise from the margins that are less stable in the document-based budget application. Romer–Romer find that durable consumption falls much more than nondurables and services, that both nonresidential and residential fixed investment decline, and that net exports improve, with exports rising and imports falling. In the support-indexed GMM estimates, the domestic investment channel remains the clearest common element: nonresidential fixed investment falls strongly and persistently. The remaining margins are less aligned. Durables fall early but later rebound, nondurables and services move little, residential investment turns positive at longer horizons, and the external-sector responses go in the opposite direction of the original RR pattern. I therefore treat these panels as component diagnostics rather than as the central mechanism evidence.

Tables OA.9–OA.11 decompose the mechanism outcomes using the same additive object as Paper Table 6. The pattern is consistent with the main-text interpretation. The medium-run decline in inflation and the large fall in investment are mainly associated with income-tax components, especially individual income taxes. Consumption moves in the same direction but less sharply. The nonstandard

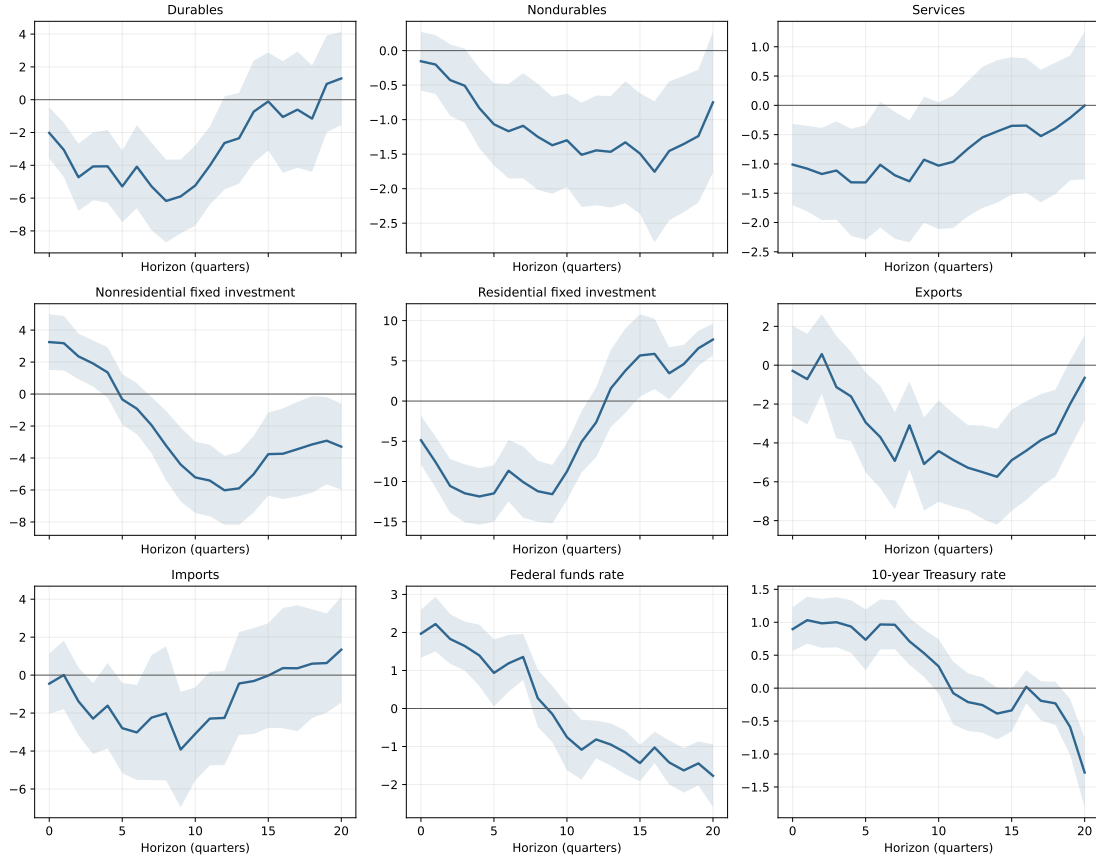


Figure OA.7: Additional RR-Style Component Responses Under the Baseline GMM Estimator

Note: The figure reports affine $p = 1$ endpoint responses using the same budget-measure probabilities, quarterly tax shocks, endpoint denominator, and HAC inference as Paper Figure 6. Components are log quantity indexes, so responses are percent changes. The panels correspond to Panels B–D in Figure 14 of Romer & Romer (2010).

consumption/excise residual continues to behave differently from a clean indirect-tax object, reinforcing the decision to treat the UK–France VAT evidence as the sharper consumption-tax contrast.

Table OA.9: Composition and Additive Inflation Contributions to the U.S. Budget Endpoint

Instrument	Event share	Gross tax share	Mean p	Inflation $h = 4$	Inflation $h = 8$	Inflation $h = 12$
Individual income	30.6%	37.0%	0.72	0.23 (0.39)	-0.34 (0.47)	-0.82 (0.41)
Corporate income	32.2%	22.7%	0.71	0.20 (0.19)	-0.03 (0.18)	-0.40 (0.23)
Clean standard excise	8.9%	11.2%	0.52	-0.05 (0.07)	-0.13 (0.13)	-0.14 (0.11)
Nonstandard consumption/excise	11.1%	15.5%	0.53	-0.31 (0.24)	-0.23 (0.18)	0.11 (0.12)
Other low-support taxes	17.2%	13.6%	0.50	0.15 (0.12)	0.16 (0.18)	0.10 (0.15)

Note: Entries are additive contributions to the aggregate affine $p = 1$ budget endpoint; parentheses report fully stacked HAC standard errors that account for fitted control projections and the common endpoint denominator. The rows use the aggregate support index and common endpoint denominator and therefore sum to the aggregate response up to rounding.

Table OA.10: Composition and Additive Consumption Contributions to the U.S. Budget Endpoint

Instrument	Event share	Gross tax share	Mean p	Consumption $h = 4$	Consumption $h = 8$	Consumption $h = 12$
Individual income	30.6%	37.0%	0.72	-1.95 (0.71)	-2.20 (0.94)	-1.78 (1.02)
Corporate income	32.2%	22.7%	0.71	-0.63 (0.39)	-0.76 (0.51)	-0.47 (0.44)
Clean standard excise	8.9%	11.2%	0.52	0.01 (0.12)	-0.11 (0.13)	-0.05 (0.08)
Nonstandard consumption/excise	11.1%	15.5%	0.53	0.78 (0.49)	0.95 (0.63)	1.07 (0.60)
Other low-support taxes	17.2%	13.6%	0.50	0.38 (0.26)	0.25 (0.21)	0.01 (0.19)

Note: Entries are additive contributions to the aggregate affine $p = 1$ budget endpoint; parentheses report fully stacked HAC standard errors that account for fitted control projections and the common endpoint denominator. The rows use the aggregate support index and common endpoint denominator and therefore sum to the aggregate response up to rounding.

Table OA.11: Composition and Additive Investment Contributions to the U.S. Budget Endpoint

Instrument	Event share	Gross tax share	Mean p	Investment $h = 4$	Investment $h = 8$	Investment $h = 12$
Individual income	30.6%	37.0%	0.72	-4.50 (1.84)	-6.71 (3.08)	-6.76 (2.36)
Corporate income	32.2%	22.7%	0.71	-1.78 (1.12)	-2.59 (1.66)	-1.78 (1.02)
Clean standard excise	8.9%	11.2%	0.52	-0.04 (0.58)	-0.40 (0.41)	-0.07 (0.31)
Nonstandard consumption/excise	11.1%	15.5%	0.53	1.69 (0.93)	2.63 (1.72)	2.61 (1.38)
Other low-support taxes	17.2%	13.6%	0.50	1.10 (0.96)	0.58 (0.59)	-0.05 (0.56)

Note: Entries are additive contributions to the aggregate affine $p = 1$ budget endpoint; parentheses report fully stacked HAC standard errors that account for fitted control projections and the common endpoint denominator. The rows use the aggregate support index and common endpoint denominator and therefore sum to the aggregate response up to rounding.

OA.4.3 Complete Instrument-Specific Response Paths

Paper Appendix D.3 distinguishes the joint instrument-specific exercise from additive contributions and states its stronger moment, exclusion and rank restrictions. Figures OA.8–OA.10 report the same estimates over horizons zero to twenty. Each panel holds the other contemporaneous tax groups fixed under those restrictions; the panels are not additive contributions.

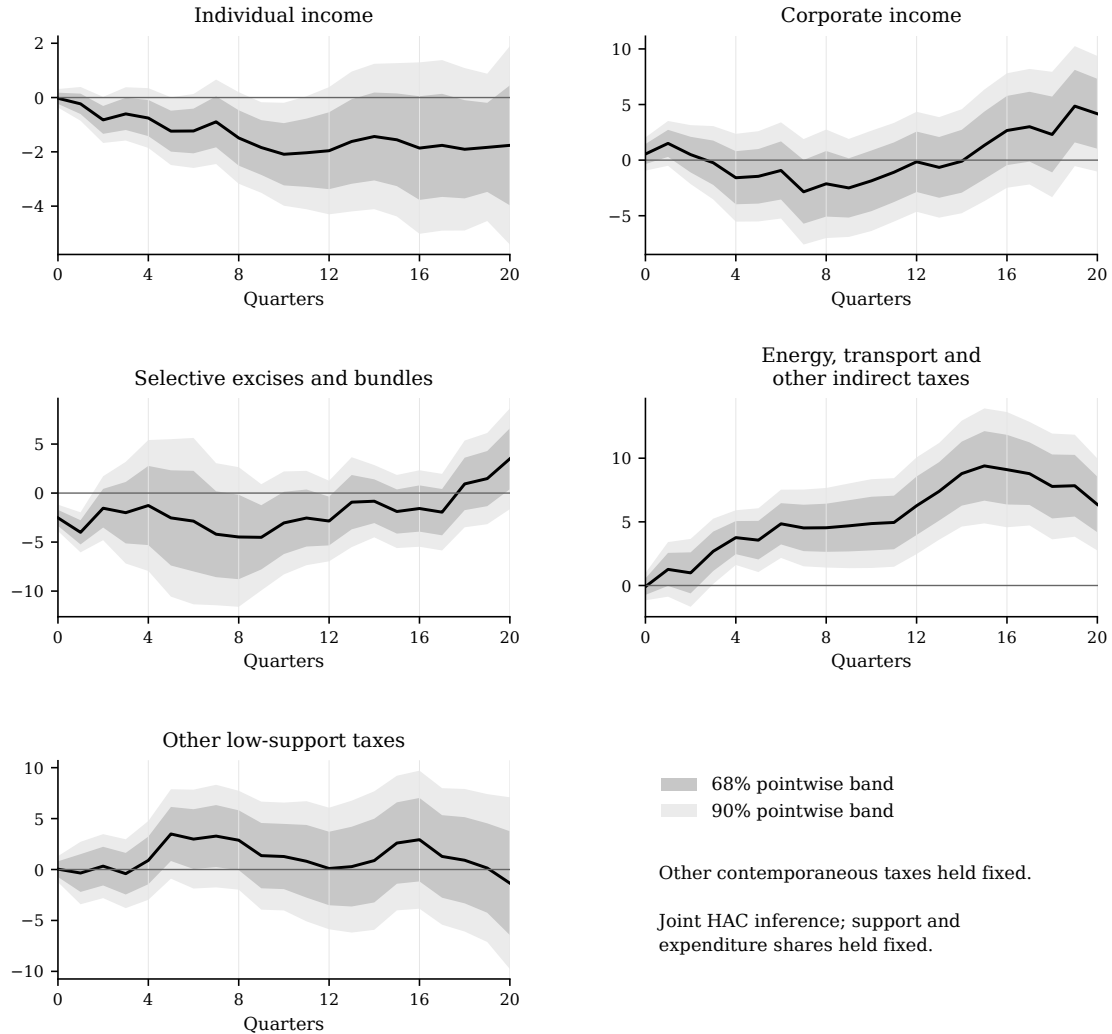


Figure OA.8: Instrument-Specific GDP Responses

Note: Percent GDP response to a proposed revenue increase equivalent to one percent of GDP in the indicated tax group. Shaded regions are pointwise 68% and 90% fully stacked HAC intervals, including the control projections and all numerator and own- and cross-tax moments. The sample, support construction, zero-filling and no-action conventions match Paper Appendix D.3. Inference conditions on constructed support and is not weak-identification-robust.

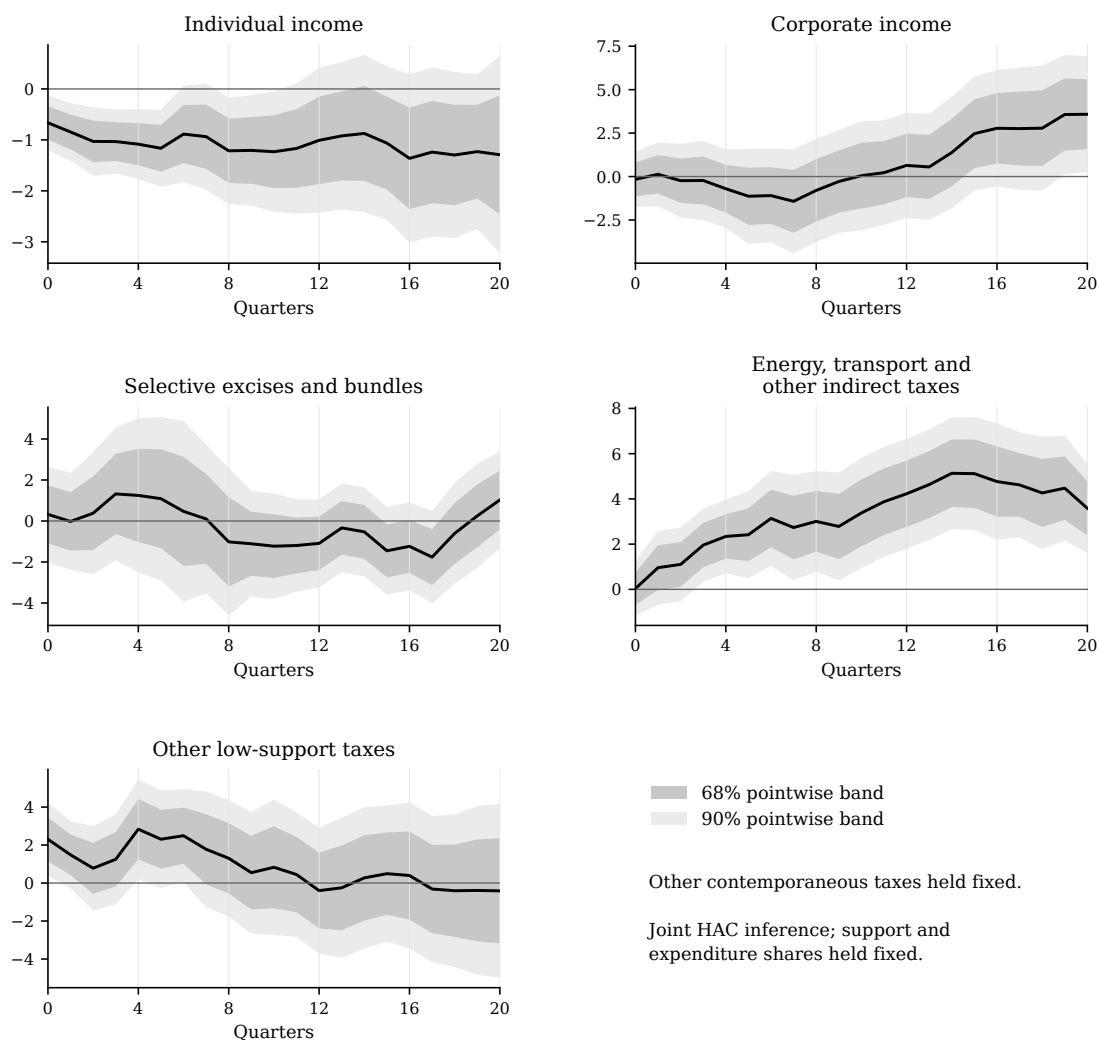


Figure OA.9: Instrument-Specific Consumption Responses in GDP Units

Note: Approximate percentage-point GDP contribution of the consumption response to a proposed revenue increase equivalent to one percent of GDP in the indicated tax group. The percent consumption response is multiplied by the fixed sample-mean nominal consumption share, 0.6250; this is not a change in the consumption/GDP ratio. Bands and maintained restrictions are as in Figure OA.8. The expenditure share is held fixed for inference, and estimates need not sum to the aggregate consumption response.

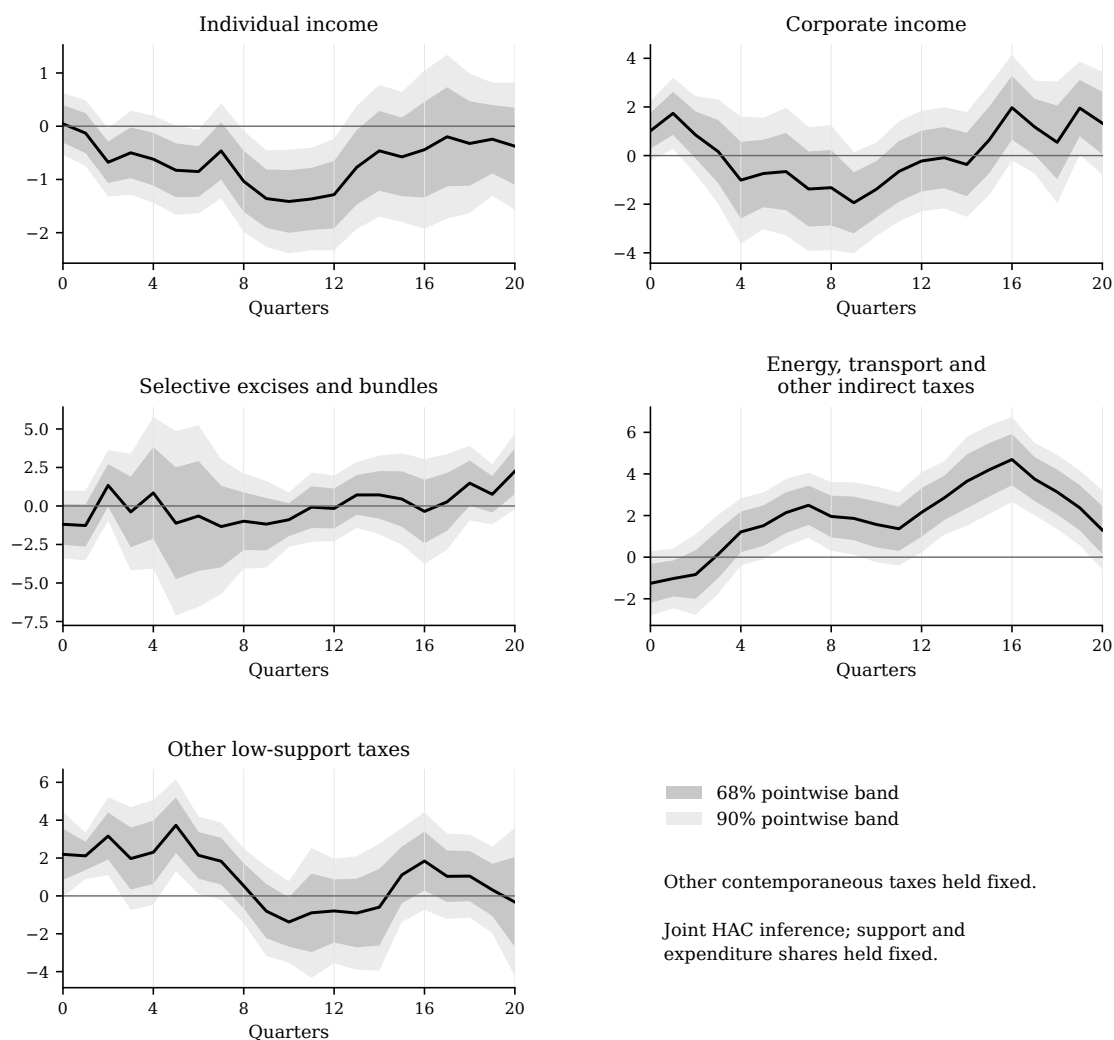


Figure OA.10: Instrument-Specific Investment Responses in GDP Units

Note: Approximate percentage-point GDP contribution of the investment response to a proposed revenue increase equivalent to one percent of GDP in the indicated tax group. The percent investment response is multiplied by the fixed sample-mean nominal investment share, 0.1746; this is not a change in the investment/GDP ratio. Bands and maintained restrictions are as in Figure OA.8. The expenditure share is held fixed for inference, and estimates need not sum to the aggregate investment response.

OA.5 A Structural Tax-Wedge Benchmark

Structural benchmark and pure tax wedges. This subsection uses the calibrated model of Trabandt and Uhlig (2011) as an external benchmark for the expenditure composition of the empirical responses. The model is not estimated, its parameters are not chosen to match the empirical impulse responses, and the exercise does not compare empirical multipliers with steady-state Laffer curves. Its narrower purpose is to ask whether distinct tax wedges generate different expenditure compositions within a common structural environment.

The baseline experiments are permanent, initially unanticipated increases in capital-income, labor-income, and consumption tax rates. Each is scaled to raise mechanical ex-ante tax liabilities by one percent of initial GDP, holding the initial tax base fixed; this requires rate increases of 5.58, 1.61, and 1.66 percentage points, respectively, in the baseline U.S. calibration. Government purchases and net imports remain fixed, so the residual fiscal adjustment is equivalent to an implicit lump-sum transfer rather than additional spending. The experiments isolate tax wedges under a common fiscal closure rather than reproduce complete historical fiscal packages.

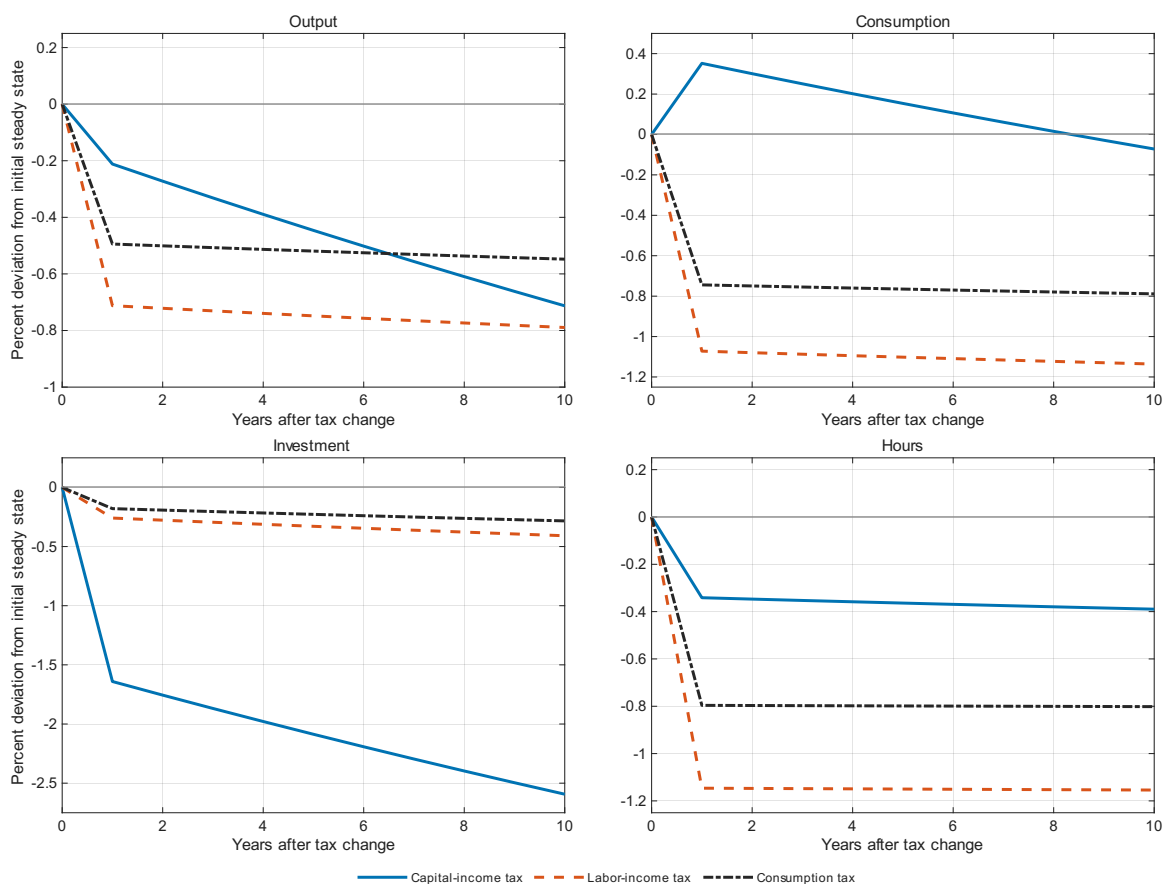


Figure OA.11: Trabandt–Uhlig Benchmark Transitions after Alternative Tax Increases

Note: The figure reports annual perfect-foresight transitions in the baseline U.S. calibration of Trabandt & Uhlig (2011) following permanent, initially unanticipated tax-rate increases. Each increase is scaled to raise mechanical ex-ante tax liabilities by one percent of initial GDP, holding the initial tax base fixed. Government purchases and net imports remain fixed; the residual fiscal adjustment is equivalent to an implicit lump-sum transfer. Responses are percent deviations from the initial steady state. Model parameters and responses are not fitted to the empirical estimates. The model contains no nominal or monetary-policy block.

Pure capital-income taxation generates a contraction concentrated in investment, while pure labor-income and consumption taxation operate predominantly through consumption and hours. Consumption initially rises under the capital-income-tax experiment and first falls below its initial level in model year 9, so this pure wedge is not a direct counterpart to the empirical U.S. response. The pure experiments instead identify the margins underlying the joint-income exercise below.

Mapping the empirical income-tax category into model wedges. The empirical administrative categories do not uniquely identify the labor- and capital-income content of the U.S. proposals. I therefore parameterize a joint-income-tax experiment by the capital-income share of its mechanical revenue increase,

$$\alpha = \frac{\Delta R_k^{\text{mech}}}{\Delta R_n^{\text{mech}} + \Delta R_k^{\text{mech}}}, \quad (\text{OA.3})$$

holding the total mechanical ex-ante increase fixed at one percent of initial GDP. Because the empirical policy measures are expressed in revenue units, the main comparison uses a symmetric mechanical-revenue normalization, $\alpha = 0.5$: labor- and capital-income taxation each contribute mechanically 0.5 percent of initial GDP. In the baseline U.S. calibration, this requires $\Delta\tau^n = 0.806$ percentage points and $\Delta\tau^k = 2.781$ percentage points. This is a normalization, not an estimate of the proposals' true structural composition; the baseline tax rates and the original steady state remain unchanged.

At model year 2, the symmetric mechanical-revenue experiment gives output, consumption, and investment responses of -0.492 , -0.394 , and -0.995 percent. Investment therefore responds about 2.53 times as much as consumption in proportional terms. In GDP units, however, the corresponding consumption and investment contributions are -0.236 and -0.256 percentage points and are nearly balanced. Table OA.13 uses this contribution accounting; the exercise is not calibrated to reproduce the empirical magnitudes.

Composition sensitivity. Figure OA.12 varies α while holding total mechanical income-tax revenue fixed. In proportional-response terms, investment becomes more responsive than consumption at $\alpha_2^* \simeq 0.289$. GDP contributions become balanced only much closer to the 50/50 revenue benchmark, as the values above show. The figure therefore makes the interpretive limit transparent: both the proportional responses and their GDP contributions depend on a labor-capital composition that is not identified by the empirical administrative categories.

Table OA.12: Alternative Normalizations for the Joint Income-Tax Experiment

Normalization	α	$\Delta\tau^n$	$\Delta\tau^k$	Year-2 $ I / C $
Equal percentage points	0.225	1.250	1.250	0.766
Equal proportional changes	0.273	1.172	1.519	0.937
Equal mechanical revenue	0.500	0.806	2.781	2.526

Note: α is the capital-income share of the total mechanical ex-ante revenue increase. Tax-rate changes are in percentage points. Each normalization raises total mechanical income-tax liabilities by one percent of initial GDP, holding initial bases fixed. The final column reports the absolute investment-to-consumption response ratio in model year 2.

Equal percentage-point movements place 22.5 percent of the mechanical revenue increase on capital income and produce the larger proportional response in consumption at year 2. Equal proportional tax-rate movements place 27.3 percent on capital income and yield nearly balanced proportional responses. Equal mechanical revenue contributions produce the larger proportional response in investment, while the two GDP contributions are nearly equal. These alternatives reinforce that statutory symmetry and revenue symmetry are distinct economic normalizations.

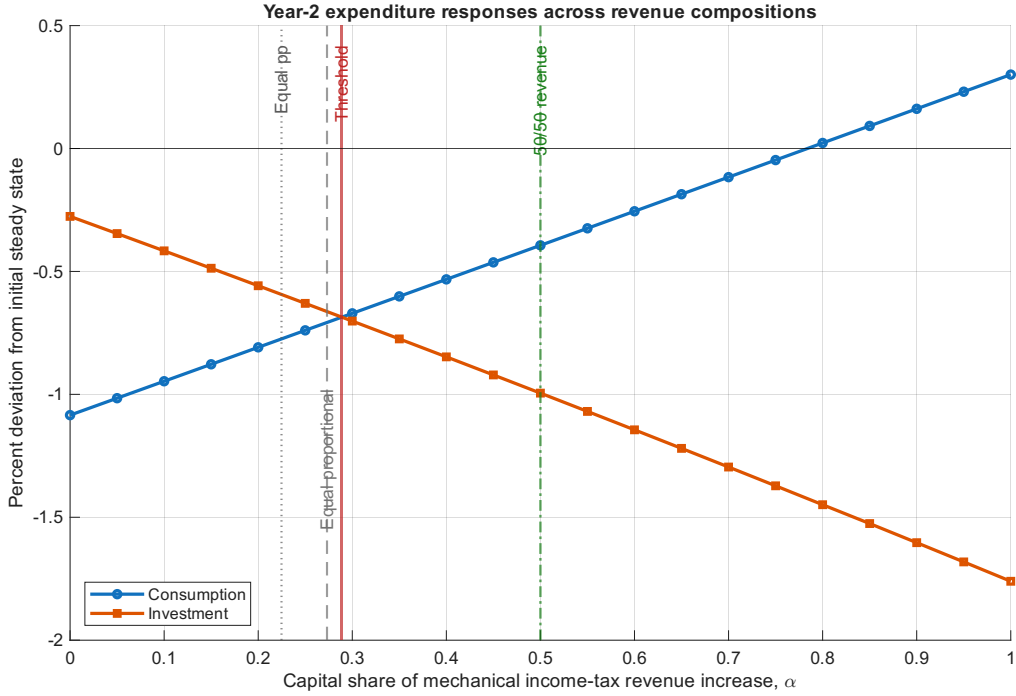


Figure OA.12: Income-Tax Revenue Composition and Year-2 Expenditure Responses

Note: The figure reports model-year-2 consumption and investment responses in the baseline U.S. calibration as the capital-income share α of a one-percent-of-initial-GDP mechanical income-tax revenue increase varies. Responses are percent deviations from the initial steady state. Equal percentage-point movements imply $\alpha \simeq 0.225$, equal proportional tax-rate movements imply $\alpha \simeq 0.273$, and equal mechanical revenue contributions imply $\alpha = 0.5$. The vertical threshold, $\alpha_2^* \simeq 0.289$, is where the absolute consumption and investment responses are equal. The model is not fitted to the empirical responses.

Consumption taxes across calibrations. Consumption accounts for most of the model output contraction under consumption taxation in each calibration relevant for the VAT application. At model year 2, the contributions to output, consumption, and investment are $(-0.504, -0.454, -0.050)$ percentage points of initial GDP in the United States. The corresponding triples are $(-0.491, -0.444, -0.047)$ in France and $(-0.459, -0.420, -0.039)$ in the United Kingdom. Within the calibrated model, the consumption-centered contribution pattern is therefore not specific to the European calibrations. This does not establish that the empirical UK–France VAT magnitudes transport quantitatively to the United States.

Table OA.13: Empirical and Structural Expenditure Contributions in GDP Units

Experiment	Output	Consumption	Investment
<i>Panel A. Empirical endpoint responses—quarter 8</i>			
U.S., income-tax-heavy	−1.874	−1.170	−1.136
UK–France, VAT	−4.547	−2.466	−1.303
<i>Panel B. Trabandt–Uhlig simulations—model year 2</i>			
U.S., joint income tax (50/50 revenue)	−0.492	−0.236	−0.256
U.S., consumption tax	−0.504	−0.454	−0.050
France, consumption tax	−0.491	−0.444	−0.047
UK, consumption tax	−0.459	−0.420	−0.039

Note: Entries are percentage-point changes relative to GDP. For the U.S. empirical row, consumption and investment are first-order contributions obtained by multiplying the log responses by their full-sample average nominal expenditure shares; the output entry is the GDP log response and has the same first-order interpretation. VAT outcomes are estimated directly in percentage points of trend GDP. Model entries are exact level changes relative to initial model GDP, $100(X_h - X_0)/Y_0$. Government purchases and net imports are fixed in the model, so its consumption and investment contributions sum to output. Each model tax change raises mechanical ex-ante liabilities by one percentage point of initial GDP; the joint-income experiment allocates half to labor-income taxation and half to capital-income taxation. Quarter 8 and model year 2 represent the same elapsed horizon. The simulations are qualitative benchmarks, not fitted structural responses.

Empirical comparison in GDP units. Table OA.13 places the empirical and model expenditure responses in GDP-contribution units. At horizon 8, the approximate U.S. consumption and investment contributions are nearly equal, at -1.17 and -1.14 percentage points of GDP. Under the symmetric mechanical-revenue normalization, the joint-income-tax simulation is similarly balanced, with model contributions of -0.236 and -0.256 . This model comparison is conditional on the assumed labor–capital composition: alternative compositions generate different expenditure rankings. By contrast, private consumption makes the larger measured contribution to the empirical VAT contraction and accounts for the larger share of the model consumption-tax response. The model does not match the empirical magnitudes, but its contribution accounting provides a qualitative benchmark for how the relevant tax wedge can alter the expenditure composition of the output response. Because the empirical U.S. income-tax measures do not map uniquely into labor- and capital-income wedges, the symmetric normalization remains a transparent benchmark rather than a fitted explanation. The model contains no nominal block and cannot speak to the VAT inflation response.

Robustness and limitations. Halving the size of the symmetric mechanical-revenue joint-income experiment to a 0.5-percent-of-GDP mechanical revenue increase leaves its expenditure ranking unchanged; after rescaling, its year-2 responses differ from the baseline per-unit responses by at most 1.2 percent. The central comparison is therefore not a large-shock artifact over this range, although it remains a numerical result within the calibrated economy.

Several limits remain. The model is annual, whereas the empirical responses are quarterly; model year 2 and empirical quarter 8 are compared at the same elapsed horizon for a qualitative assessment.

Structural tax wedges do not map one-for-one into administrative narrative measures, and the labor–capital composition of the empirical income-tax proposals remains unidentified. The simulations use the model’s transfer-adjustment fiscal closure. Finally, the model has no price level, nominal rigidity, or monetary-policy block and therefore cannot speak to the VAT inflation response.

OA.6 Supplementary VAT Diagnostics

OA.6.1 Additional VAT Channel Diagnostics

Figure OA.13 reports the fuller set of expenditure-side responses. The figure keeps the same outcome transformation as the main mechanism figure: each component is expressed as a change from the pre-event quarter in percentage points of trend GDP. The main text focuses on private consumption and investment because they are the central domestic-absorption channels for a VAT shock. The additional panels show that government consumption is muted at short horizons, while net exports partly offset the contraction. Because the real component series are chain-linked and the reported set omits inventories and statistical discrepancies, their sum need not equal the GDP response; the corresponding residual is documented with the table source data.

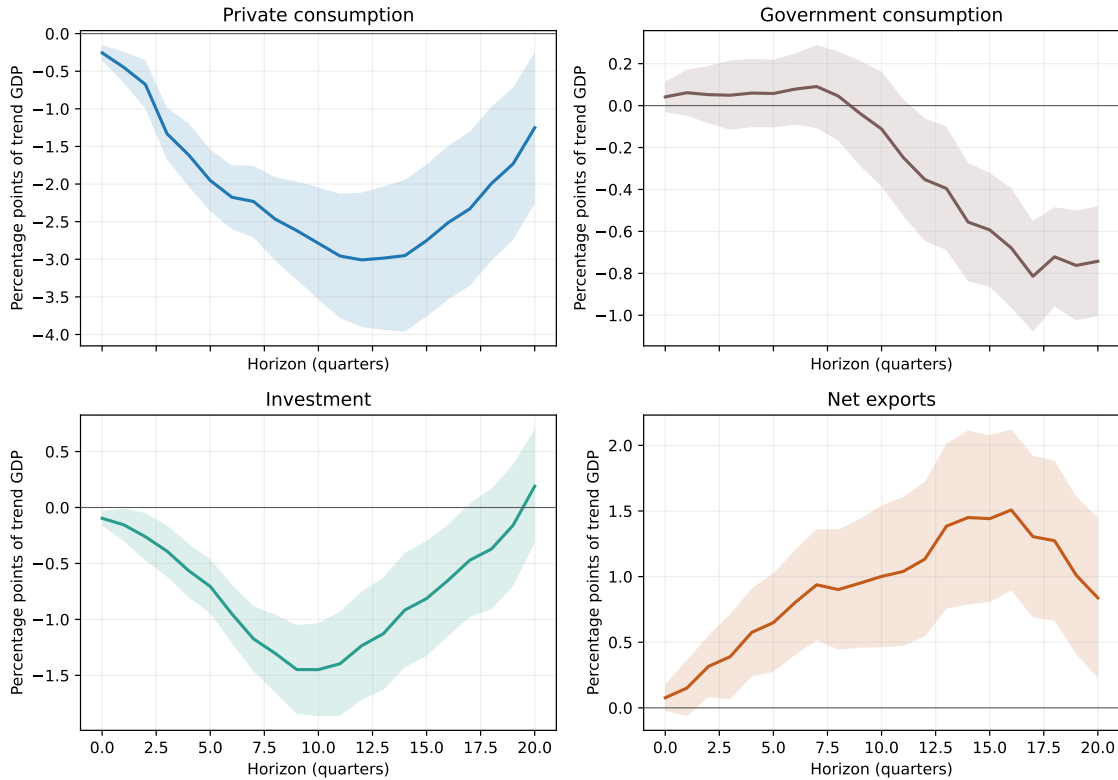


Figure OA.13: UK–France VAT: Additional Expenditure-Side Responses

Note: Outcomes are expressed as $y_{c,t+h} - y_{c,t-1}$ in percentage points of trend GDP. The figure reports the pooled VAT endpoint for private consumption, government consumption, investment, and net exports. Shaded areas are 68 percent fully stacked HAC confidence bands.

Table OA.14 reports the corresponding key-horizon estimates. Decomposing net exports shows that the external offset is mostly import compression rather than a robust export boom: exports are small and imprecise, while the import contribution is positive. Since imports enter the accounting decomposition as $-M$, a positive import contribution means that imports fall.

Table OA.14: VAT Additional Channels at Key Horizons

Outcome	$h = 4$	$h = 8$	$h = 12$
GDP deviation from trend	-2.46	-4.55	-4.04
Private consumption response	-1.61	-2.47	-3.01
Government consumption response	0.06	0.05	-0.35
Investment response	-0.56	-1.30	-1.23
Exports response	0.17	-0.37	0.16
Imports contribution ($-M$)	0.41	1.27	0.97
Net exports response	0.58	0.90	1.13
Reported component sum	-1.54	-2.82	-3.46
Accounting residual	-0.92	-1.72	-0.57

Note: GDP and expenditure components are changes from the pre-event quarter in percentage points of trend GDP. Inflation is in percentage points. Imports enter as $-M$, so a positive import contribution denotes import compression. The reported component sum is private consumption plus government consumption plus investment plus exports minus imports. The accounting residual is the GDP response less that sum; it absorbs omitted inventories and statistical discrepancies as well as chain-weighting differences.

OA.6.2 Complete Paths Across Support Constructions

The key-horizon construction evidence is reported in Paper Appendix E.1.

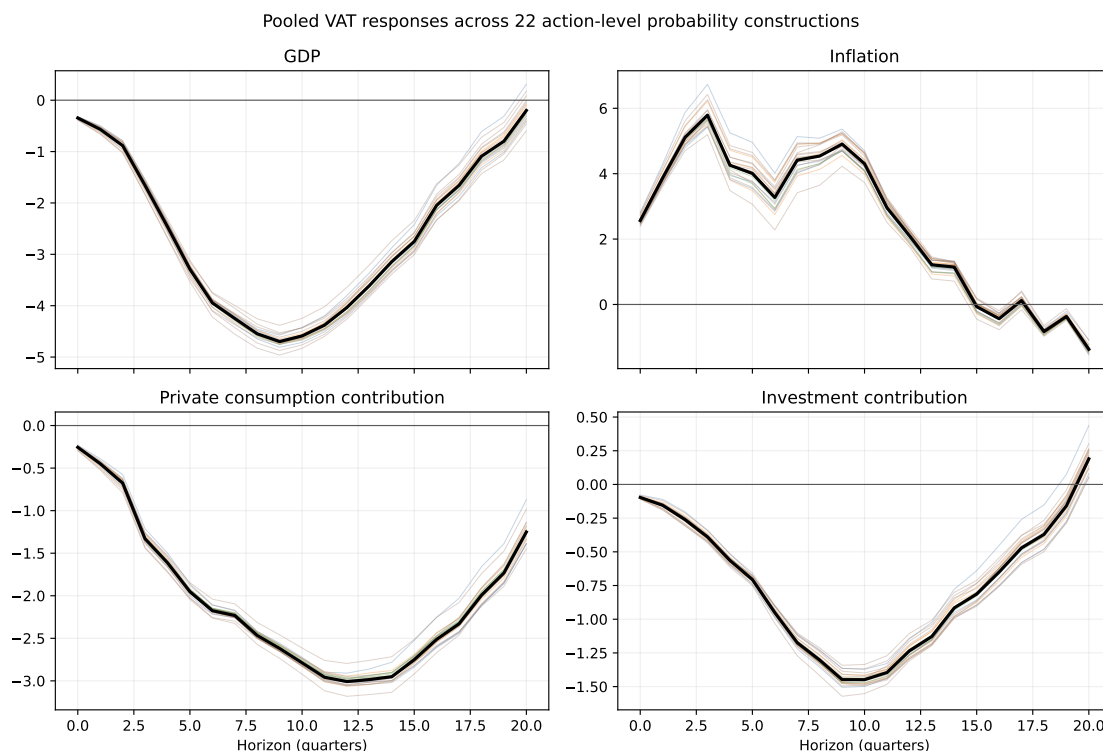


Figure OA.14: VAT Responses Across Action-Level Support Constructions

Note: The black line is the baseline ensemble and the colored lines are the 21 alternative action-level support constructions summarized in Paper Table 23. GDP, private consumption, and investment are changes from the pre-event quarter in percentage points of trend GDP; inflation is in percentage points. Every specification retains the baseline full-calendar convention, including support equal to zero in quarters without a recorded VAT action.

OA.6.3 Complete Package-Sensitivity Paths

The specifications and key-horizon estimates are reported in Paper Appendix E.2.

OA.6.4 Country-Specific VAT Diagnostics

Figure OA.16 reports the same output and price endpoints separately for France and the United Kingdom on the common comparison sample. These country-specific estimates are supporting diagnostics rather than the main estimand. They show that the pooled GDP response is not averaging two opposite country patterns: GDP falls in both countries after high-support VAT shocks. The country-specific inflation responses are less precisely estimated, but both countries display a positive short-run price response.

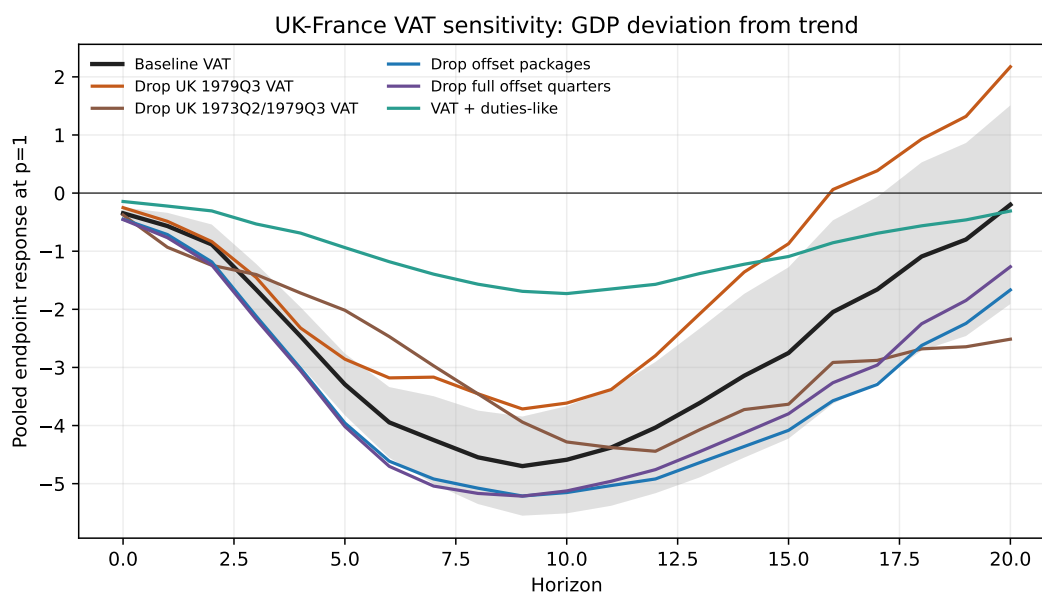


Figure OA.15: VAT GDP Response Under Package Sensitivity Checks

Note: The figure reports pooled GDP endpoint responses under the same sensitivity variants as Paper Table 24. The shaded area is the 68 percent fully stacked HAC band for the baseline VAT endpoint.

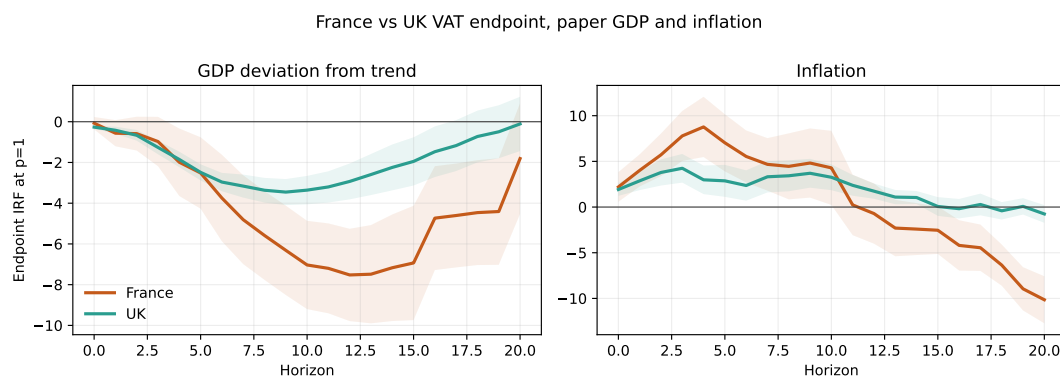


Figure OA.16: Country-Specific VAT Endpoint Responses

Note: The figure estimates the VAT endpoint separately for France and the United Kingdom on the common comparison sample, 1955Q1–2007Q4. GDP is expressed in percentage points of trend GDP; inflation is in percentage points. Responses are normalized by the country-specific linear-in- p denominator evaluated at $p = 1$. Shaded areas are 68 percent fully stacked HAC confidence bands.

Table OA.15: Country-Specific VAT Endpoint Responses at Key Horizons

Country	Outcome	$h = 4$	$h = 8$	$h = 12$
France	GDP deviation from trend	-1.99	-5.58	-7.52
France	Inflation	8.77	4.45	-0.71
United Kingdom	GDP deviation from trend	-1.85	-3.36	-2.93
United Kingdom	Inflation	2.98	3.42	1.74

Note: Entries are country-specific high-support VAT endpoints. GDP is in percentage points of trend GDP and inflation is in percentage points.

References

- Cloyne, J. (2013). Discretionary Tax Changes and the Macroeconomy: New Narrative Evidence from the United Kingdom. *American Economic Review*, 103(4), 1507–1528. <https://doi.org/10.1257/aer.103.4.1507>
- Romer & Romer. (2010). The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks. *American Economic Review*, 100(3), 763–801. <https://doi.org/10.1257/aer.100.3.763>
- Trabandt, M., & Uhlig, H. (2011). The Laffer curve revisited. *Journal of Monetary Economics*, 58(4), 305–327. <https://doi.org/10.1016/j.jmoneco.2011.07.003>