

Probabilistic Narrative Identification: Evidence from Tax Policy*

Tom Pessó[†]

October 12, 2026

Job Market Paper
[\[Latest version here\]](#)

Abstract

This paper develops a probabilistic approach to narrative identification. Traditional narrative studies retain policy actions classified as exogenous and exclude the remainder from identification. I instead assign each action a probability of exogeneity, allowing the broader policy record to remain informative. Variation in this probability allows policy actions to be reweighted so that the endogenous component cancels, recovering the component associated with exogenous policy. I construct these probabilities through a replicable mapping from policy documents using large language models. Applied to the canonical tax-policy sample of Romer and Romer (2010), the framework yields an estimated medium-run GDP contraction about two-thirds smaller than under binary selection. I then extend the analysis to indirect taxation by constructing two new datasets of U.S. budget and French VAT proposals. VAT illustrates how probabilistic identification makes narrative analysis feasible when binary selection leaves too little credible identifying variation. Together, the applications reveal contrasting transmission patterns: U.S. tax increases lower output and inflation, with similar consumption and investment contributions to the GDP decline, whereas VAT increases initially raise prices and contract output, with consumption contributing more than investment.

Keywords: Narrative identification; fiscal policy; tax shocks; causal inference.

JEL Codes: C21; C26; E62; E32.

*I am deeply grateful to Geert Mesters for invaluable guidance and to Mark Watson for insightful comments. I also thank Davide Debortoli and Albert Marcet for helpful comments and advice, and the participants of the CREi internal seminar for helpful feedback. This research was supported by the Scientific Computing Core Facility (MELIS-UPF), which generously provided access to computational resources. All errors are my own.

[†]Universitat Pompeu Fabra (UPF), Barcelona School of Economics (BSE). Contact: tom.pesso@upf.edu

Introduction

Identifying the macroeconomic effects of fiscal policy is difficult because policy actions respond to current and anticipated economic conditions. A large literature addresses this identification problem using restrictions on the dynamics of macroeconomic variables or external identifying information, including instruments and narrative evidence (Ramey, 2016). Narrative identification exploits contemporaneous policy documents, including policymakers’ stated motivations, to isolate policy actions that are plausibly exogenous to macroeconomic conditions. This approach has been applied across a wide range of settings, including monetary policy, taxation, government spending, and oil-market disruptions (Friedman & Schwartz, 1963; Hamilton, 1985; Mertens & Ravn, 2012; Ramey, 2011; Romer & Romer, 1989, 2010). Its implementation requires researchers to judge whether the documentary evidence is sufficiently convincing to treat each policy action as exogenous to contemporaneous and anticipated macroeconomic conditions.

In practice, however, researchers often face incomplete, ambiguous, or conflicting evidence in the documentary record, leaving the case for exogeneity uncertain and open to interpretation. A tax action may, for example, pursue a long-run reform objective while also reflecting concerns about current economic conditions. Nevertheless, traditional narrative designs rely on a binary inclusion decision, retaining policy actions for which the evidence is judged sufficiently convincing to support an exogenous interpretation and excluding the remainder from identification. A demanding standard retains only a few strongly supported policy actions, leaving limited identifying variation; a more permissive standard broadens coverage but includes actions for which the case for exogeneity is less clear.

The paper’s main contribution is to develop probabilistic narrative identification, a framework that uses uncertainty about narrative classifications as information for identification. I complement the framework with a document-based procedure that uses large language models to recover policymakers’ stated motivations and maps these motivations into narrative-support probabilities for individual policy actions. The framework is particularly valuable for studying value-added tax (VAT), where infrequent reforms and mixed motivations can leave little credible identifying variation under binary selection.

The identification argument starts from a distinction familiar from traditional narrative identification: I represent each policy action as belonging either to an exogenous type or to a policy-response type reflecting systematic responses to economic or fiscal conditions.¹ However, the researcher does not observe which type an action belongs to. The documentary record provides a probability p that the action belongs to the exogenous type, which I refer to as *narrative support for exogeneity*. Intermediate values of p therefore represent uncertainty about exogeneity. To give p this probability interpretation, I assume that narrative support is calibrated in population: among policies assigned support p , the fraction belonging to the exogenous type is p . Traditional narrative identification emerges as a special case when a cutoff on p determines which actions are retained and the retained tax variation is assumed to provide orthogonal policy variation.

To see how this probability object provides identifying information, consider its implications for observed outcome–tax comovement. Reflecting the two unobserved policy types, this comovement combines two components that cannot be observed separately: one associated with orthogonal policy

¹For a causal interpretation, policy variation within the exogenous type is assumed orthogonal to other drivers of the outcome. Traditional binary narrative identification imposes an analogous restriction on the policy variation retained after classification.

variation and the other with systematic policy responses. Traditional narrative identification isolates the first component by treating policy actions classified as exogenous as a source of orthogonal policy variation. The probabilistic framework instead uses the additional identifying variation in p to describe how the two components enter observed outcome–tax comovement: the orthogonal component receives weight p and the systematic component weight $1 - p$. Different values of p therefore reveal different weighted averages of the same two unknown components of outcome–tax comovement.²

The key idea is to combine these weighted averages in a way that eliminates the systematic-policy contribution while preserving the orthogonal-policy contribution. With two distinct values of p , for example, we have two weighted averages of two unknown comovement components, allowing their contributions to be separated through appropriate reweighting. This reweighting also admits an instrumental-variables representation: multiplying each observed tax change by a weighting function $w(p)$ produces an instrument for that tax change. This connects probabilistic narrative identification to external-instrument methods widely used in applied macroeconomics (Stock & Watson, 2018). Section 1 formalizes these restrictions and the identification argument.

The identification and estimation framework treats p as observed; in practice, I augment the collected policy dataset with a measure of narrative support constructed from documentary evidence. I develop a procedure that, for each policy, locates relevant passages in contemporaneous documents, recovers policymakers’ stated motivations, and uses large language models to construct numerical representations of the meaning conveyed by those motivations. Language-model representations are particularly useful in this setting because narrative classifications often hinge on subtle contextual differences in stated motivations and on ambiguities that simple lexical measures may not capture. Their scalability also makes it feasible to conduct narrative analysis at the level of individual policy proposals, matching the unit of measurement to the level at which policies are stated and motivated. A fitted classifier maps these representations into a measure of narrative support, whose probability interpretation rests on the population-calibration assumption stated above.³ The document-based measurement procedure as a whole concentrates researcher judgment in a limited set of explicit choices, including the motivation categories, language model, and calibration sample, rather than repeated case-by-case classification decisions.⁴ The procedure is designed to make the measurement replicable while preserving an auditable trail from the underlying passages and stated motivations to each resulting support score.⁵

I apply the framework to tax policy, retaining the full available policy record and using uncertainty about narrative classification as identifying information. In the Romer–Romer benchmark, probabilistic identification yields a substantially more muted medium-run GDP point estimate. I then broaden coverage to individual U.S. budget proposals, including consumption and excise measures alongside income taxes. The aggregate contraction is concentrated in income-tax proposals, while indirect taxes make no similarly clear contractionary contribution. I therefore turn to VAT as a more clearly defined indirect-tax instrument. VAT reforms are infrequent and often have mixed motivations, so a

²This interpretation requires within-type moment stability: the average outcome–tax moment within each policy type does not vary with narrative support.

³The baseline classifier is a multinomial logit for policy-motivation categories.

⁴I assess how the empirical results change across alternative ways of measuring p , including varying the open-weight large language model used to construct the representations. Because these measurements may share distortions, I also use the sensitivity analysis of Armstrong and Kolesár (2021) to examine how large such distortions would need to be to overturn the empirical conclusions.

⁵Exact numerical replication may still depend on the hardware used, as computations can differ slightly across machines when generating outputs from fixed open-weight large language models.

demanding binary selection rule can leave little credible identifying variation. Probabilistic narrative identification keeps these reforms informative without requiring a sharp inclusion decision.

The constructed probabilities capture economically meaningful variation in narrative support for exogeneity. I compare them with the narrative classifications of Romer and Romer (2010) and Cloyne (2013), which provide external benchmarks for the document-based measure. In both samples, policies classified as exogenous receive higher narrative support on average. In the Romer–Romer sample, however, policies classified as exogenous receive average support of only 0.60, indicating substantial residual uncertainty among the policies retained for identification.

Retaining the full available policy record and using uncertainty about narrative classification as identifying information yields a substantially smaller estimated medium-run GDP contraction than under traditional binary selection. In the Romer–Romer sample, for a tax increase equivalent to one percent of GDP, the probabilistic specification yields a GDP point estimate of -0.78 percent after two years, compared with -2.25 percent under the RR binary-selection benchmark using the same macroeconomic outcomes and lag controls.⁶ When continuous narrative support is instead converted into binary inclusion rules, one cutoff produces an estimate close to the RR benchmark, but the magnitude of the medium-run contraction remains sensitive to the chosen threshold.⁷

I then combine a new granular policy dataset of U.S. budget tax proposals with probabilistic narrative identification to study how tax composition shapes macroeconomic transmission. Different tax instruments need not transmit through the same macroeconomic margins (Mertens & Ravn, 2013). To preserve this instrument-level detail, I collect individual proposals from the annual federal budgets, recording their projected revenue effects, implementation dates, and tax categories.⁸ Relative to narrative series organized around major historical episodes, this broadens coverage of federal tax-policy variation while preserving the composition of the reported proposals.⁹ I then augment this policy dataset with a narrative-support measure for each proposal. Using the augmented dataset, I apply the probabilistic framework to study individual tax instruments where a demanding binary selection rule can leave little credible identifying variation. Together, the broader record and the probabilistic framework allow me to assess which instruments contribute to the aggregate response, including taxes outside the personal and corporate income-tax margins emphasized in earlier narrative work.

I find that the contractionary aggregate response is concentrated in individual and corporate income-tax proposals, while indirect taxes make no similarly clear contribution. Across this broader policy record, the estimated response to a proposed tax increase equivalent to one percent of GDP is a sizable output contraction accompanied by sustained medium-run disinflation: after two years, GDP declines by 1.87 percent and inflation by 0.57 percentage points. Consumption and investment make approximately equal contributions to the GDP decline after two years. Negative medium-run

⁶At two years, the smaller response mainly reflects a larger fitted squared-tax normalization. The comparison also uses different control-fitting populations, as detailed in Section 5.

⁷This exercise applies traditional binary selection to the continuous support measure. The probabilistic framework uses support directly and requires no inclusion cutoff. Section 5 discusses this comparison.

⁸The budget documents record tax proposals together with their contemporaneous rationales and reported implementation dates, following the spirit of Cloyne (2013). The analysis focuses on whether policy actions respond systematically to macroeconomic conditions through their stated motivations. Anticipation between announcement and implementation is a distinct identification issue (Leeper et al., 2013; Ramey, 2011) and is not addressed here.

⁹Unlike Romer and Romer (2010), which record major enacted tax changes, the budget dataset records administration proposals at the individual-policy level. The two sources can overlap when a budget proposal is subsequently enacted, but their units of observation and stages of the policy process differ.

GDP point estimates for income-tax increases also emerge when other tax instruments are held fixed, under stronger identifying restrictions. The U.S. federal record, however, combines selective excises and levies affecting production as well as consumption. This leaves open how a broad consumption-tax instrument shapes macroeconomic transmission.

To study consumption-tax transmission more directly, I construct a new long-run policy dataset of French VAT proposals from 1960 to 2024 and combine it with UK VAT changes from Cloyne (2013). VAT reforms are infrequent and often motivated by several competing objectives, so a demanding binary selection rule can leave little credible identifying variation. A key empirical payoff of the probabilistic framework is to retain these mixed-motive reforms without requiring a sharp inclusion decision. VAT provides a more clearly defined indirect-tax instrument, with consumer prices and household spending central to its transmission (Bachmann et al., 2026). The pooled UK–France estimates show that VAT increases raise prices initially and contract output. Private consumption makes a larger contribution than investment to the GDP decline after two years.

Taken together, the two applications highlight the importance of interpreting aggregate responses in light of their tax composition. In the U.S. budget application, where the contractionary aggregate contribution is concentrated in individual and corporate income-tax proposals, consumption and investment make approximately balanced contributions to GDP after two years and inflation declines. For VAT, the contraction is instead more concentrated in private consumption and accompanied by an initial increase in prices. These patterns are consistent with the different margins affected by income and consumption taxation.¹⁰

Related Literature

Narrative and Macroeconomic Identification. This paper relates first to the literature identifying policy shocks from time-series restrictions and the historical record. Structural VARs recover shocks through timing, sign, or other model-based restrictions (Blanchard & Perotti, 2002; Sims, 1980; Uhlig, 2005), while external-instrument methods use variables that are correlated with the target policy shock and orthogonal to other structural innovations (Ramey, 2016; Stock & Watson, 2018). Narrative approaches provide a distinct source of identifying information: contemporaneous records are used to construct policy-shock measures that can enter estimation directly or serve as external instruments. Foundational applications include monetary policy (Friedman & Schwartz, 1963; Romer & Romer, 1989, 2004), oil and government spending shocks (Hamilton, 1985; Ramey, 2011; Ramey & Shapiro, 1998), and tax changes (Cloyne, 2013; Romer & Romer, 2010). Recent work continues to extend narrative identification to new settings, including global shipping disruptions and historical tariff policy (Bini, 2025; den Besten et al., 2026; Känzig & Raghavan, 2026). Related work combines sparse narrative shocks with additional time-series information (Barnichon & Mesters, 2025). This paper extends traditional narrative identification by using uncertainty about narrative classifications as information for identification.

Latent Components, Moment Recovery, and Constructed Instruments. The econometric structure relates most directly to work that uses observation-specific probabilities of belonging to un-

¹⁰Simulations from the calibrated model of Trabandt and Uhlig (2011) illustrate this contrast. An equal mechanical-revenue split between labor- and capital-income tax changes yields approximately balanced consumption and investment contributions, whereas a consumption-tax change produces a more consumption-centered contraction.

observed groups to recover group-specific moments. In mixture models with varying concentrations, these mixing probabilities permit recovery of component moments through weights that need not be positive (Doronin & Maiboroda, 2014; Liubashenko & Maiboroda, 2015; Maiboroda et al., 2013). The same intuition applies here: groups of policy actions with different narrative support have different proportions of exogenous and policy-response types. Under calibration and within-type moment stability, observed outcome–tax moments can therefore be combined so that the policy-response-type moment cancels while the exogenous-type moment is preserved. Here, however, narrative support is measured from documentary evidence, and its probability interpretation rests on maintained calibration. The resulting reweighting also admits an IV representation, with narrative support determining the weights applied to the observed policy change. This connects more broadly to methods that construct instruments from observed variation under additional restrictions, for instance restrictions on heteroskedasticity (Lewbel, 2012; Rigobon, 2003; Rigobon & Sack, 2004). The same moment-recovery argument yields both the constructed instrument and a conditional restriction that I exploit through a finite set of unconditional moments (Chamberlain, 1987; Newey, 1990).

Text and AI in Economic Analysis. A growing literature uses text to measure economically relevant objects that are not directly observed in conventional data (Ash & Hansen, 2023; Gentzkow et al., 2019). Earlier applications use corporate disclosures to measure product-market similarity (Hoberg & Phillips, 2016) and central-bank communications to measure monetary-policy information and deliberation (Hansen & McMahon, 2016; Hansen et al., 2018). More recent work uses language models to recover structured information from economic documents, including labels, summaries, rationales, and other features that would traditionally require human coding (Clayton et al., 2025; Jamilov, 2025; Martell, 2025). Another strand uses model representations and probabilities as statistical signals for classification, uncertainty assessment, and the construction of economic measures (Buckmann & Hill, 2025; Chen et al., 2026; Hartley, 2025). This paper combines these two uses. I first recover policymakers’ stated motivations from contemporaneous fiscal documents and then use statistical representations of those motivations to construct calibrated narrative support for exogeneity. The resulting measurement strategy makes systematic narrative analysis feasible at the level of individual policy actions, while preserving differences in their stated motivations and policy composition. These probabilities then enter the identifying moments of the probabilistic narrative framework.

AI-Generated Measurements and Imperfect Moments. Because narrative support is generated from text, the paper relates to recent work on inference with variables constructed from unstructured data (Battaglia et al., 2024; Carlson & Dell, 2025; Ludwig et al., 2026). This literature emphasizes that generated measurements may contain systematic error and develops corrections when validation data provide information about the target variable. Direct validation is more demanding here because the target is a continuous probability of latent exogeneity, for which existing binary expert classifications do not provide observed counterparts. Beyond assessing construction sensitivity, I draw on corrected-moment methods (Evdokimov & Zeleneev, 2022) and approximate-moment sensitivity analysis (Armstrong & Kolesár, 2021) to distinguish correction under an explicit measurement model from robustness to departures from the identifying moments.

Tax Transmission and Composition. The empirical applications connect to work on the macroeconomic effects of tax changes. Narrative evidence shows that tax increases are contractionary (Cloyne, 2013; Mertens & Ravn, 2012; Romer & Romer, 2010), while subsequent work emphasizes differences across personal, corporate, and consumption taxes (Mertens & Ravn, 2013; Nguyen et al., 2021; Trabandt & Uhlig, 2011). The U.S. application extends narrative tax evidence to a granular record of routine federal tax proposals and uses measure-level information to characterize the tax composition of the aggregate high-support response. This differs from Mertens and Ravn (2013), who separately identify personal and corporate income-tax shocks to estimate tax-specific causal responses. The VAT application also relates to evidence on the identification and measurement of consumption-tax shocks (Riera-Crichton et al., 2016), VAT pass-through (Benzarti et al., 2020), and household consumption responses (Bachmann et al., 2026; Cashin & Unayama, 2016). The new French narrative proposal series, combined with UK VAT changes, links these indirect-tax margins to aggregate output, prices, and consumption over a long historical sample.

The remainder of the paper is organized as follows. Section 1 develops the identification framework. Sections 2 and 3 present the document-based measurement procedure and its implementation for tax policy. Section 4 describes aggregation, estimation, inference, and measurement-error sensitivity. Section 5 presents the external benchmarks and the Romer–Romer comparison. Sections 6 and 7 study U.S. budget tax proposals and VAT reforms in the United Kingdom and France, respectively. Section 8 concludes.

1 A Probabilistic Model of Narrative Identification

The objective is to estimate the dynamic causal effects of discretionary tax-policy actions on macroeconomic outcomes using narrative evidence from policy documents. Let $\Delta\tau_t$ denote the aggregate tax-policy change recorded in period t , and let y_{t+h} denote an outcome of interest at horizon $h \geq 0$. Consider the structural relation

$$y_{t+h} = \beta_h \Delta\tau_t + \varepsilon_{t+h},$$

where β_h is the causal response to exogenous tax variation and ε_{t+h} collects other determinants of the outcome. The identification problem arises because tax policy may respond systematically to economic and fiscal conditions that also enter ε_{t+h} . Multiplying the structural relation by $\Delta\tau_t$ gives

$$\mathbb{E}(y_{t+h} \Delta\tau_t) = \beta_h \mathbb{E}(\Delta\tau_t^2) + \mathbb{E}(\Delta\tau_t \varepsilon_{t+h}).$$

The last term need not be zero when tax changes respond, for example, to economic activity, inflation, fiscal pressure, or current spending decisions. Identification therefore requires separating outcome–tax comovement associated with orthogonal policy variation from that associated with systematic policy responses.

Traditional narrative approaches use contemporaneous documentary evidence to make this distinction, retaining policy actions judged sufficiently orthogonal for identification and excluding the remaining variation. This strategy has proved influential in empirical work on tax policy, as in Romer and Romer (2010), Mertens and Ravn (2013), and Cloyne (2013). The documentary record, however, often leaves the case for exogeneity uncertain. Policy actions may combine several motivations, and the available documents may provide stronger support for an exogenous interpretation for some ac-

tions than for others.¹¹ The framework below instead treats variation in documentary support for exogeneity as identifying information. I first formalize traditional binary narrative identification as a benchmark and then show how probabilistic narrative support can be used to separate the moments associated with orthogonal and systematic policy variation.

1.1 Binary Narrative Identification

Traditional narrative identification maps documentary evidence into a binary inclusion decision: policy actions are retained for identification when the documentary record is judged to provide sufficiently convincing evidence that they are unrelated to current or anticipated macroeconomic conditions; the remaining actions are excluded. Let $\mathbb{1}\{\Delta\tau_t \text{ is retained}\}$ denote this decision.

The binary narrative decision can be represented schematically as a threshold rule on documentary support. If p_t indexes this support,

$$\mathbb{1}\{\Delta\tau_t \text{ is retained}\} = \mathbb{1}\{p_t \geq \bar{p}\},$$

for some cutoff $\bar{p} \in (0, 1)$. Identification additionally requires the retained tax changes to satisfy

$$\mathbb{E}[\varepsilon_{t+h}\Delta\tau_t \mid p_t \geq \bar{p}] = 0, \quad h \geq 0.$$

Binary narrative identification therefore combines a judgment about the documentary evidence with an orthogonality restriction on the retained policy variation.

This representation highlights two implications of binary narrative identification. First, policy actions below the cutoff do not contribute identifying information, even when the documentary record provides substantial support for an exogenous interpretation that falls short of the cutoff. Second, once the cutoff is crossed, residual uncertainty is set aside: retained actions are treated as providing orthogonal policy variation. The probabilistic framework instead retains this variation in narrative support as identifying information.¹²

1.2 Probabilistic Narrative Support and Policy-Response Moments

Traditional narrative identification relies on a binary inclusion decision that, for purposes of identification, treats the policy type as resolved by the documentary evidence. A key element of the framework is that the policy type remains unobserved, so that documentary evidence only provides probabilistic information about type membership. Let $u_t \in \{0, 1\}$ denote the latent type of tax change $\Delta\tau_t$, with $u_t = 1$ denoting the exogenous type and $u_t = 0$ the policy-response type reflecting systematic responses to current or anticipated economic or fiscal conditions.¹³ The remaining uncertainty about policy type is summarized by narrative support $p_t \in [0, 1]$.

Narrative support becomes informative for the macroeconomic relationship of interest through the outcome–tax moment. At a given value of p_t , this moment depends on two objects: the relative

¹¹The same issue arises for other narrative instruments. For example, geopolitical motivations for defense-news shocks may provide strong evidence against a domestic stabilization response without implying exogeneity with probability one (Ramey & Zubairy, 2018).

¹²Appendix A.6 examines the trade-off between bias and sampling uncertainty under binary selection and probabilistic identification, holding the policy series fixed.

¹³Appendix A.3 considers the stronger alternative in which a single tax-policy action contains both exogenous and policy-response components and p_t calibrates the latent exogenous share relevant for the outcome–tax moment.

incidence of the exogenous and policy-response types, and the outcome–tax comovement within each type. The restrictions below specify how narrative support relates to each.

Narrative-support calibration. To give p_t a probability interpretation, we impose population calibration:¹⁴

$$\mathbb{E}(u_t \mid p_t = p) = \Pr(u_t = 1 \mid p_t = p) = p. \quad (1)$$

Calibration therefore determines how the two policy types enter the conditional outcome–tax moment: at support p , the exogenous type has probability p and the policy-response type probability $1 - p$. However, the type-specific outcome–tax moments, $\mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = j, p_t = p)$, remain general functions of p , for $j \in \{0, 1\}$.

Within-type moment stability. Recovering the two type-specific moments therefore requires restricting how they vary with narrative support. To do so, I impose within-type moment stability: for each horizon $h \geq 0$ and all p in the support of p_t ,¹⁵

$$\mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = j, p_t = p) = \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = j), \quad j \in \{0, 1\}. \quad (2)$$

In words, this restriction means that among actions of the same policy type, the average outcome–tax moment should not vary systematically with narrative support. However, this does not mean that all actions within a type need to have identical moments. A policy type may contain different subtypes with different outcome–tax moments: for example, demand-management and spending-driven tax actions may both belong to the policy-response type while having different average moments. Such heterogeneity is compatible with the restriction, provided that it does not make the average outcome–tax moment for the policy type vary with p .

When a policy type contains different subtypes, within-type moment stability constrains both the moments within subtypes and their relative incidence.¹⁶ A natural, slightly stronger sufficient condition is that the average outcome–tax moment within each subtype be stable across p and that changes in subtype composition not alter the average moment for the policy type.

Support-indexed moments. Calibration and within-type moment stability can now be combined to give a simple representation of the observed outcome–tax moment:¹⁷

$$\mathbb{E}(y_{t+h}\Delta\tau_t \mid p_t = p) = p \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1) + (1 - p) \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0). \quad (3)$$

Equation (3) is the central identifying relation. Narrative support determines the relative weight placed on two fixed moments: the exogenous-type moment and the policy-response-type moment. Different values of p therefore provide different weighted averages of these same two moments. At

¹⁴More generally, one could allow $\Pr(u_t = 1 \mid p_t = p) = g(p)$. The baseline imposes $g(p) = p$, giving p a direct probability interpretation rather than merely an ordinal ranking.

¹⁵State dependence can be accommodated by imposing calibration and within-type moment stability conditional on the observed state, provided sufficient support variation remains within states. Appendix A.2 shows how averaging the recovered moments within the exogenous type preserves the baseline causal interpretation.

¹⁶Appendix A.1.1 formally derives this decomposition, separating variation in outcome–tax moments within subtypes from changes in subtype composition across narrative-support levels.

¹⁷By iterated expectations over u_t , equation (1) supplies the weights p and $1 - p$, while equation (2) removes conditioning on p_t from the two type-specific moments.

$p = 1$, the systematic-policy component receives zero weight and the support-indexed moment equals the exogenous-type moment. Section 1.3 shows how variation in p can recover this moment without requiring $p = 1$ to be observed.

Exogenous-type orthogonality. Finally, as in traditional narrative identification, giving the exogenous-type moment a causal interpretation requires an orthogonality restriction. Here, I impose this restriction on tax variation of the latent exogenous type:

$$\mathbb{E}(\Delta\tau_t \varepsilon_{t+h} \mid u_t = 1) = 0, \quad h \geq 0.$$

This is the macroeconomic orthogonality restriction on the latent policy variation that narrative evidence is intended to isolate.¹⁸

Normalization. The causal response per unit of tax change follows by normalizing the exogenous-type outcome–tax moment by the corresponding squared-tax moment. The orthogonality restriction gives $\mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1) = \beta_h \mathbb{E}(\Delta\tau_t^2 \mid u_t = 1)$. In the baseline, I recover the required squared-tax moment by imposing the analogous within-type stability restriction on $\Delta\tau_t^2$ and applying the same support-indexed argument. Provided this moment is positive, the ratio of the exogenous-type outcome–tax and squared-tax moments identifies β_h .

1.3 Identification from Variation in Narrative Support

Variation in narrative support provides different weighted averages of the exogenous- and policy-response-type moments, allowing the two to be separated. The case with two support values makes the recovery of the orthogonal-policy moment transparent and leads naturally to an instrumental-variables representation.

Two support values. Suppose p_t takes two values, p_L and p_H , with $p_H > p_L$. Equation (3) then implies

$$\begin{aligned} \mathbb{E}(y_{t+h}\Delta\tau_t \mid p_t = p_L) &= p_L \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1) + (1 - p_L) \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0), \\ \mathbb{E}(y_{t+h}\Delta\tau_t \mid p_t = p_H) &= p_H \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1) + (1 - p_H) \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0). \end{aligned}$$

These are two weighted averages of the same type-specific moments. Solving for the exogenous-type moment gives

$$\frac{1 - p_L}{p_H - p_L} \mathbb{E}(y_{t+h}\Delta\tau_t \mid p_t = p_H) - \frac{1 - p_H}{p_H - p_L} \mathbb{E}(y_{t+h}\Delta\tau_t \mid p_t = p_L) = \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1).$$

The weights preserve the exogenous-type moment while making the policy-response-type moment cancel. Importantly, neither support group needs to consist only of exogenous policy actions: identification comes from their different probabilities of belonging to the exogenous type. The high-support moment receives positive weight, while the low-support moment is subtracted to remove the systematic-policy contribution.

¹⁸Calibration, within-type moment stability, and exogenous-type orthogonality are maintained identifying restrictions. They are not directly testable because policy type is unobserved; orthogonality additionally involves the structural disturbance ε_{t+h} .

From two values to general support. When p_t takes many values, replace the two group weights by an observation weight $w(p_t)$ satisfying

$$\mathbb{E}[w(p_t)p_t] = 1, \quad \mathbb{E}[w(p_t)(1 - p_t)] = 0. \quad (4)$$

The first condition gives the orthogonal-policy component total weight one, while the second gives the systematic-policy component total weight zero. Applying these conditions to equation (3) gives

$$\mathbb{E}[w(p_t)y_{t+h}\Delta\tau_t] = \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1).$$

The resulting weights preserve the moment associated with orthogonal policy variation and remove the moment associated with systematic policy responses, just as in the two-value case. This argument is closely related to estimators for mixtures with varying concentrations, which use known variation in mixing probabilities and signed weights to recover component-specific moments and regressions (Doronin & Maiboroda, 2014; Liubashenko & Maiboroda, 2015; Maiboroda et al., 2013).

Instrumental-variables interpretation. Applying the same weighting argument to the squared tax change yields $\mathbb{E}[w(p_t)\Delta\tau_t^2] = \mathbb{E}(\Delta\tau_t^2 \mid u_t = 1)$, under the corresponding moment-stability restriction. The causal response can therefore be written as

$$\beta_h = \frac{\mathbb{E}[w(p_t)y_{t+h}\Delta\tau_t]}{\mathbb{E}[w(p_t)\Delta\tau_t^2]} = \frac{\mathbb{E}[(w(p_t)\Delta\tau_t)y_{t+h}]}{\mathbb{E}[(w(p_t)\Delta\tau_t)\Delta\tau_t]}.$$

This is the familiar just-identified IV ratio, with the reweighted tax change $w(p_t)\Delta\tau_t$ serving as an instrument for $\Delta\tau_t$. The weighting representation therefore provides a direct link between probabilistic narrative identification and external-instrument methods.

This representation also allows the same identification strategy to be used in other IV specifications. Under the corresponding relevance and orthogonality conditions, $w(p_t)\Delta\tau_t$ can enter IV local projections or serve as an external instrument in a structural VAR (Stock & Watson, 2018).

The recovery conditions in equation (4) do not uniquely determine $w(\cdot)$ when narrative support takes more than two values. Different weighting rules can therefore recover the same population moment. The next subsection develops the moment-condition formulation used for estimation and shows how the baseline estimator selects a particular weighting rule.

1.4 Moment Conditions and a GMM Formulation

The preceding subsection shows how variation in narrative support can be reweighted to recover the exogenous-type moment. I now express the same identification argument as moment conditions used for estimation. Equation (3) implies that the outcome–tax moment is affine in narrative support:

$$\mathbb{E}(y_{t+h}\Delta\tau_t \mid p_t = p) = \underbrace{\mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0)}_{a_h} + p \underbrace{[\mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1) - \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0)]}_{b_h}.$$

Thus, a_h is the policy-response-type moment, while the $p = 1$ endpoint $a_h + b_h$ is the exogenous-type moment identified above. Write $m_h(p; \theta_h) = a_h + b_h p$, with parameter $\theta_h = (a_h, b_h)'$.

The affine relation implies the following conditional moment restriction.¹⁹

$$\mathbb{E}[y_{t+h}\Delta\tau_t - m_h(p_t; \theta_h) \mid p_t] = 0. \quad (5)$$

Following Chamberlain (1987), multiplying equation (5) by functions of p_t yields unconditional moment conditions. For the baseline estimator, I project the conditional moment on $(1, p_t)'$, giving $\mathbb{E}[(y_{t+h}\Delta\tau_t - m_h(p_t; \theta_h))(1, p_t)'] = 0$. Provided narrative support varies sufficiently, these two moments identify a_h and b_h . The resulting just-identified GMM estimator is numerically equivalent to the OLS projection of $y_{t+h}\Delta\tau_t$ on a constant and p_t . Viewed through the weighting argument of the preceding subsection, this estimator selects a particular weighting rule that recovers the $p = 1$ endpoint.²⁰ Applying the same construction to $\Delta\tau_t^2$ recovers the exogenous-type squared-tax moment, so the ratio of the two $p = 1$ moments yields β_h .

The baseline moments do not exhaust the implications of equation (5). For any admissible instrument function $\phi(p_t)$, the conditional restriction implies $\mathbb{E}[(y_{t+h}\Delta\tau_t - m_h(p_t; \theta_h))\phi(p_t)] = 0$. The choice $\phi(p_t) = (1, p_t)'$ is sufficient to identify the two affine coefficients but need not use the conditional restriction efficiently. Richer or optimally chosen instrument functions can in principle improve precision (Chamberlain, 1987; Newey, 1990).²¹ I retain $(1, p_t)'$ as a parsimonious implementation that yields an explicit endpoint-weighting rule and provides a common finite moment system for the measurement-error and sensitivity analyses below. This choice is made for transparency and tractability rather than efficiency.

The baseline identifying restrictions also determine how the outcome–tax moment varies with narrative support: calibration and within-type moment stability deliver the linear relation in p derived above. Relaxing these identifying assumptions generally allows the support-indexed moment to vary nonlinearly with p . Such departures can be accommodated by allowing $m_h(p; \theta_h)$ to take a more general form, while retaining the conditional moment formulation above.²² Appendix A.5 characterizes the identifying restrictions associated with departures from the linear benchmark.

1.5 Measurement Error in Narrative Support

The identification argument above treats narrative support p_t as observed. In practice, the language-model procedure developed below constructs a generated measure $\hat{p}_t$ from policy documents. Discrepancies between $\hat{p}_t$ and p_t can shift the support-indexed moments used for identification, particularly when they reflect systematic features of the model, prompt, or document representation, or are related to variables entering the estimating equations. Existing approaches address this problem by combining generated measurements with trusted validation data to account for measurement error in downstream

¹⁹In the empirical applications, I impose the corresponding affine restrictions directly on the outcome–tax and squared-tax moments formed from population residuals using the control projections in Section 4.2. These restrictions are maintained for the residualized moments and do not follow mechanically from analogous restrictions on the unadjusted moments.

²⁰For the affine projection on $(1, p_t)'$, the corresponding population weight is $w(p_t) = 1 + \frac{(1 - \mathbb{E}[p_t])(p_t - \mathbb{E}[p_t])}{\text{Var}(p_t)}$, which satisfies the recovery conditions in equation (4).

²¹Conditional restrictions can also be exploited through larger collections of unconditional moments; see Carrasco and Florens (2000). Expanding the instrument set, however, introduces a finite-sample trade-off rather than guaranteeing better performance (Donald et al., 2009). I do not pursue instrument selection or efficient estimation here.

²²In the empirical implementation, outcomes and tax changes are residualized on the controls included in each macroeconomic specification. Section 4.2 gives the corresponding two-step sample estimator, while Appendix C records the application-specific samples and develops the corresponding estimation and inference procedures.

inference (Battaglia et al., 2024; Carlson & Dell, 2025; Ludwig et al., 2026).²³ These approaches are more difficult to apply here because the relevant target is a probability of latent exogeneity, for which trusted measurements are unavailable. Existing expert classifications provide a useful external benchmark for whether $\hat{p}_t$ aligns with established binary narrative judgments, as examined in Section 5, but they do not directly identify p_t or the measurement-error distribution of $\hat{p}_t$. The analysis described below therefore focuses on measurement error in the baseline moment system obtained from the projection on $(1, p_t)'$.

In the absence of trusted measurements of latent support, I first assess how sensitive the results are to alternative constructions of $\hat{p}_t$. These alternatives vary each step of the measurement procedure, including the open-weight language models used to construct narrative support. Without further assumptions, the resulting dispersion measures sensitivity to measurement choices, but does not reveal how far the constructed probabilities are from latent support: baseline confidence bands capture sampling uncertainty conditional on $\hat{p}_t$, but do not incorporate uncertainty from its measurement. Under the additional assumption that the alternative constructions provide classical repeated measurements of latent support, their differences can inform a correction to the baseline projected moments. I examine this benchmark using corrected projected moments in the spirit of Evdokimov and Zeleneev (2022).

Comparisons across alternative constructions cannot reveal systematic measurement errors common to all of them. I therefore complement these exercises by asking how large departures from the maintained identifying moments would need to be to overturn the empirical conclusions, following Armstrong and Kolesár (2021). The dispersion observed across alternative constructions provides reference magnitudes for these departures, rather than bounds on remaining measurement error.²⁴

2 Measuring Narrative Support for Exogeneity

The empirical task is to translate the qualitative evidence associated with each policy action into action-level narrative support p_i . I index individual policy actions by i ; Subsection 4.1 aggregates their scores into the quarterly p_t used in estimation. I construct p_i in three steps: recover the stated motivations for the policy action from the relevant documentary record, measure how those motivations map into economically defined narrative categories, and aggregate the resulting evidence into a policy-level support probability. Section 3 specifies the choices used to implement this construction in the tax applications. Figure 1 summarizes the construction.²⁵

A policy action may have several stated motivations, some supporting an exogenous interpretation and others indicating a response to economic or fiscal conditions. The measurement construction preserves this mixed evidence rather than resolving it into a binary classification.

2.1 From Stated Motivations to Support Probabilities

I begin with the economically substantive step, mapping stated motivations into p_i , before explaining how those motivations are recovered from documents. A policy action can have several stated moti-

²³Battaglia et al. (2024) also develop analytical and joint-estimation corrections when the upstream error process is known, estimable from external information, or identified through a sufficiently specified measurement model. These approaches likewise require information beyond the generated model outputs themselves.

²⁴Section 4.3 describes the implementation, while Appendices C.7 and C.8 provide the assumptions and derivations.

²⁵Online Appendix Figure OA.1 presents the full action-level implementation and subsequent quarterly aggregation.

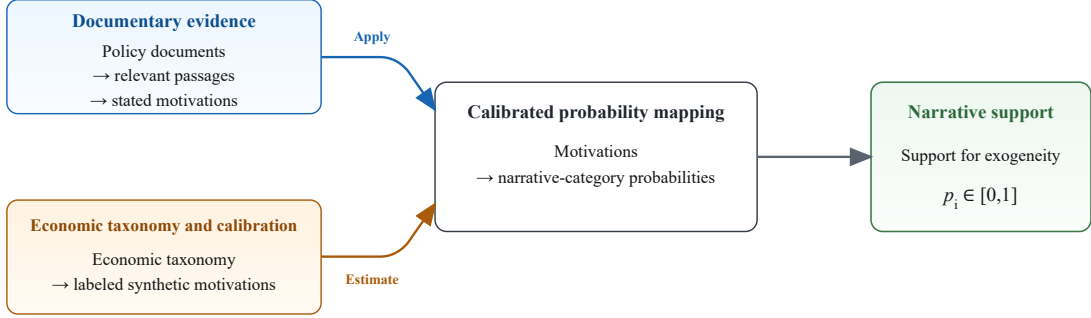


Figure 1: Construction of narrative support. Documentary evidence supplies policy-specific stated motivations. The economic taxonomy supplies labeled synthetic motivations for a calibrated probability mapping. Applying that mapping to documentary motivations yields category probabilities and action-level narrative support for exogeneity, p_i .

vations; I map each motivation separately into probability mass over a predetermined set of economic categories and then aggregate the resulting support to the policy level. Mapping stated motivations into support requires a well-defined set of economic categories against which the textual signal can be interpreted (Ash & Hansen, 2023; Gentzkow et al., 2019). Let $\mathcal{G}$ denote these categories and $\mathcal{G}_{\text{exo}} \subset \mathcal{G}$ the subset whose stated rationales support an exogenous interpretation. Section 2.2 defines the tax-specific taxonomy used in the applications. Given these categories, the empirical task is to map each stated motivation M into probability mass over $\mathcal{G}$,

$$\left(\hat{P}_c(M)\right)_{c \in \mathcal{G}}, \quad \hat{P}_c(M) \geq 0, \quad \sum_{c \in \mathcal{G}} \hat{P}_c(M) = 1.$$

Language models are useful for this task because stated policy motivations can express subtle or overlapping economic rationales that are difficult to capture with fixed dictionaries or surface-level textual features. In the baseline implementation, a multinomial logit maps language-model representations of stated motivations into probabilities over the narrative categories. I estimate this mapping using synthetic motivations generated with assigned category labels, then apply it to motivations recovered from policy documents. The economic taxonomy therefore determines the distinctions to be measured, while the calibrated statistical model determines how strongly a stated motivation loads on each category. Section 3.1 describes the calibration samples, language-model representations, and statistical specification used in the applications.²⁶

The support attached to motivation M is the probability mass assigned to the categories supporting the exogenous type:

$$p(M) = \sum_{c \in \mathcal{G}_{\text{exo}}} \hat{P}_c(M).$$

For a policy action with several stated motivations, the baseline support p_i averages their motivation-level probabilities. This aggregation preserves mixed documentary evidence at the policy level: mo-

²⁶Online Appendix OA.1.2.5 reports an alternative construction based directly on token-probability signals.

tivations associated with systematic policy responses reduce support for the exogenous type, while motivations associated with more orthogonal policy objectives increase it. Under the maintained calibration interpretation, p_i is support for assigning an action to the exogenous type, not a literal exogenous fraction of its tax change.

2.2 Tax-Policy Taxonomy and Illustration

The tax applications illustrate this construction. I use the narrative taxonomy developed by Romer and Romer (2010) and refined by Cloyne (2013). Its categories distinguish stated responses to economic or fiscal conditions from motivations that support an exogenous interpretation.

The Cloyne taxonomy decomposes the policy-response component into economically distinct channels. Demand management and supply stimulus correspond primarily to responses to activity or inflation; short-run deficit reduction responds to fiscal pressure; and spending-driven changes respond to current expenditure decisions. Each category therefore corresponds to a different component of the policy rule and to a potentially different source of bias. Long-run, ideological, external, and deficit-consolidation motives instead provide support for assignment to the exogenous type. Figure 2 reports the detailed category definitions and examples.

Group	Sub category	Explanation and examples
Endogenous (N)	1. Demand management (DM)	- Targeting the aggregate level of demand e.g. to boost investment, consumption, growth, or curb inflation. - Specific help to households and individuals by stimulating disposable income. - Dealing with a balance of payments crisis via demand.
	2. Supply stimulus (SS)	- Certain help for businesses during a downturn (e.g. NIC cut). - Short term sector support (e.g. targeted tax cuts for a sector).
	3. Deficit reduction/BoP crisis (DR)	Direct measures to deal with a budget or external deficit contemporaneously caused.
	4. Spending-driven (SD)	Taxes which fund specific spending commitments.
Exogenous (X)	1. LR performance (LR)	- Measures to improve competitiveness, productivity, efficiency and long-run growth (but not taken to offset a shock). - Simplification and deregulation measures. - Long-term support for business or sectors of the economy.
	2. Ideological (IL)	- Long-term social or political goals, independent of their effect on performance and not to offset current shocks. - Some anti-avoidance measures (where no other motive is given).
	3. External (ET)	Court rulings and enforcement of directives.
	4. Deficit consolidation (DC)	- Measures to lower inherited deficit for reasons of economic philosophy or to offset current actions in the future. - Does not include actions forced on the government, or decisions contemporaneous motivated by a current shock.

Figure 2: Narrative Classification of Tax Changes, Reproduced from Cloyne (2013)

The 1977 French VAT reform fixes ideas and illustrates the ambiguity that motivates this probabilistic mapping. The reform was quantitatively important: it reduced the standard VAT rate from 20 percent to 17.6 percent, merged it with the intermediate rate, and carried a reported fiscal impact of 8.5 billion francs, about 0.5 percent of GDP. It therefore provides substantial policy variation for studying the macroeconomic effects of taxation, while its contemporaneous documentary record contains both short-run stabilization and structural-reform motives. The retrieved documentary evidence

is condensed into the following three stated motivations:

- **Mitigate the rise in prices after the end of the price freeze:** *the measure aims to limit the inflationary rebound likely to occur at the end of the price-freeze period by reducing the tax burden on consumption.*
- **Complete a structural reform:** *the measure is the final step in a reform undertaken since 1969 to merge the standard VAT rate with the intermediate rate and reduce the number of VAT rates.*
- **Fight inflation:** *the reduction of the rate is part of the broader government programme to combat inflation by easing indirect taxation.*

The first and third motivations concern current inflation and point toward demand management, while the second describes a longer-run reform of the tax system. The construction retains this mixed evidence through probabilities over the narrative categories. Subsection 3.1 reports the fitted mapping for this example and the resulting support score.

2.3 Recovering Stated Motivations from Documents

The preceding subsection takes stated motivations as given. Recovering them is difficult when the policy rationale occupies only a small part of a long documentary record. The procedure locates passages relevant to the policy action within its contemporaneous documents. Retrieval therefore restricts the evidence available for motivation extraction to passages likely to contain the stated rationale for the policy action; it does not itself classify the action. Semantic-distance retrieval is the baseline implementation, but other auditable ways of locating the relevant evidence, including manually indexed documents or model-generated document maps, could be used instead.

The retrieved passages are then converted into a short set of stated motivations. The language model is instructed to recover only rationales supported by those passages and not to infer motivations absent from the documentary record. This step has two roles. First, it isolates the substantive reasons for the policy action from the retrieved text. Second, it expresses those reasons in a compact and comparable form that can be mapped into category probabilities and manually audited against the source passages. For the 1977 VAT reform, this stage produces the three motivations quoted above.

Traceability is maintained by preserving the evidence chain behind each score. Starting from p_i , one can recover the motivation-level category probabilities, the extracted motivations on which they were computed, and the passages in the original documents supporting those motivations. The calibration and aggregation choices used along this chain are recorded in the replication material. This permits manual audit of the documentary basis of the score. The archived materials permit reproduction of the probability construction and estimation conditional on the fixed motivation sets.²⁷ Subsection 3.2 reports the embedding models, document segmentation, retrieval depth, extraction prompts, and language models used in the applications.

The next section gives the tax-policy implementation. Section 4 then turns the resulting action-level measures into quarterly inputs for estimation.

²⁷Online Appendix OA.1.1.3 documents the coverage of the original extraction records.

3 Measurement Implementation for Tax Policy

This section describes how I implement the probability mapping for tax policy, construct its documentary inputs, and assess whether the recovered motivations are grounded in the source passages.

3.1 Synthetic Calibration and Probability Construction

Following Cloyne (2013), the economic distinction underlying the taxonomy can be represented schematically as

$$\Delta\tau_i = \xi_i + r(y_i, \pi_i, b_i, g_i),$$

where y_i denotes the state of economic activity, including relevant labor-market conditions; π_i is inflation; b_i captures deficit or debt pressure; and g_i captures current spending decisions. The function $r(\cdot)$ captures systematic tax-policy responses to these conditions, while ξ_i denotes variation motivated by objectives unrelated to them.

These category distinctions also clarify the within-type stability restriction in Section 1. The scalar p_i should measure support for exogeneity rather than differences in how the measurement procedure scores distinct policy-response motives. If deficit-, demand-, supply-, and spending-driven policy changes receive systematically different support scores for reasons unrelated to exogenous membership, and these channels have different outcome-tax moments, then p_i also sorts the composition of policy-response bias. The affine slope would then combine changes in the share of exogenous actions with changes in the nature of the endogenous policy response, potentially distorting the $p = 1$ endpoint. Estimating separate probabilities for the eight motives before aggregating them into p_i helps place these channels on a common scale and makes channel-specific differences in the scores observable.

Each extracted motivation is evaluated against the definitions of these eight motives. The definitions are stated in fixed prompts, making the classification criteria explicit and reproducible. They can be used either directly to score historical motivations or to generate labeled synthetic examples, as described below.²⁸

There are two ways to use the category prompts to extract a classification signal from a language model. A direct approach examines how the model’s token-probability distribution changes when a historical motivation is added to a category prompt.²⁹ The baseline instead uses the category definitions to generate synthetic motivations with known categories and estimates a probability mapping from their language-model representations. Buckmann and Hill (2025) show that hidden states can retain classification-relevant information that is not reliably revealed by a model’s generated answer, and that simple statistical classifiers trained on these representations can match or outperform direct prompting with relatively few labeled observations. Hartley (2025) similarly treats the mapping from language-model representations to probabilities as an estimable component of economic measurement and emphasizes the value of retaining probabilistic outputs when constructing aggregate indices.

The category prompts provide a compact semantic description of what each narrative category includes and excludes. In the baseline construction, these definitions are used to sample a varied collection of motivations for each category, following approaches that generate labeled text for downstream classification (Beck et al., 2026; Ye et al., 2022). The resulting samples allow the researcher

²⁸The prompt templates are reported in Online Appendix OA.1.2.1.

²⁹Online Appendix OA.1.2.5 describes this alternative.

to inspect the motivations allocated to the different categories before estimating the probability mapping. By contrast, when classification is obtained directly from a prompt, the model’s interpretation of the category remains largely implicit and may differ across models because their training produces different representations of the same semantic distinctions.

I implement the probability mapping defined in Section 2.1 by estimating a multinomial logit of the known synthetic category labels on the hidden-state representations. Coefficient shrinkage is selected by stratified cross-validation to limit overfitting.³⁰ Applying the fitted model to a historical motivation yields probabilities over the eight categories. I sum the probabilities associated with exogenous policy variation and, when an action has several motivations, average their support as described in Section 2.1. The synthetic exercise calibrates probabilities over narrative categories. Interpreting their sum as a probability of latent exogeneity additionally requires the population calibration condition in Section 1, which the synthetic exercise cannot test because latent type is unobserved.

The synthetic calibration samples are constructed to span both clear and more ambiguous expressions of each narrative category. I generate four types of motivations. Canonical examples closely follow the category definitions, while institutional examples use the language and context typical of fiscal-policy documents. Boundary examples introduce features associated with a nearby category while preserving a clear primary motivation. Finally, mixed-motive examples include multiple stated considerations but designate one category as primary. In the latter two cases, the additional considerations remain in the text, while the primary motivation determines the label used to estimate the probability mapping.

For each of these four designs, I generate three independently seeded calibration samples. Each sample contains 70 motivations for each of the eight categories, for a total of 560 observations. This produces twelve separate calibration samples. The Romer–Romer and Cloyne observations are not used in their construction or in estimating the probability mappings and are reserved for external validation.³¹

I estimate the probability mapping separately for each of the twelve calibration samples and each of the three language-model representations. A given historical motivation therefore receives 36 probability estimates. The baseline averages across these estimates, reducing dependence on a particular synthetic sample or representation model, while their dispersion provides a measure of sensitivity to these choices. Appendix B reports holdout performance and agreement across sets and models.³² How these alternative constructions enter the measurement-sensitivity analysis is discussed below. Reducing the hidden-state representations to 32 or 64 principal components additionally preserves policy rankings, external-benchmark separation, and the main U.S. GDP and inflation responses.³³

Table 1 applies the fitted mapping to the three motivations for the 1977 VAT reform introduced in Subsection 2.2.

The first and third motivations load mainly on Demand Management and therefore contribute primarily to the policy-response categories. The second places most of its probability mass on the exogeneity-supporting Long Run and External categories. Aggregating across the three motivations yields an interior policy-level support probability: 0.45 of the estimated probability mass supports the

³⁰Online Appendix OA.1.2.3 provides the estimation details.

³¹Online Appendix OA.1.2.2 provides the generation instructions and examples; the complete synthetic samples are included in the replication materials.

³²Online Appendix OA.2.1 reports the construction audits.

³³Online Appendix OA.2.2 reports this dimensionality check.

Table 1: Calibrated Category Probabilities for the 1977 VAT Reform Example

	Category	M1	M2	M3	Overall policy prob.
Endo	Demand Management (DM)	0.76	0.00	0.81	0.55
	Spending-Driven (SD)	0.02	0.00	0.01	
	Supply Stimulus (SS)	0.03	0.00	0.00	
	Deficit Reduction (DR)	0.01	0.01	0.01	
Exo	Deficit Consolidation (DC)	0.10	0.09	0.13	0.45
	External (EXT)	0.08	0.41	0.03	
	Ideological (IL)	0.00	0.01	0.01	
	Long-Run (LR)	0.01	0.48	0.00	

Note: M1–M3 denote the price-freeze, structural-reform, and anti-inflation motivations, respectively, introduced in Section 2.2. Each motivation column sums to one across the eight categories. The shaded cell marks the highest category within each motivation. The final column reports probability mass assigned to the endogenous and exogenous blocks after averaging across motivations.

exogenous interpretation, while 0.55 supports the policy-response interpretation.³⁴

This interior score allows the reform to remain informative without classifying it as exogenous or excluding it. Under the identifying restrictions in Section 1.3, differences in narrative support can be used to reweight observed outcome–tax moments so that the policy-response-type moment cancels while the exogenous-type moment is preserved. The score is therefore an input to this reweighting across observations, not the weight itself; cancellation is a property of the combined moments, not of this reform considered in isolation.

3.2 Document Processing and Motivation Recovery

Let q_i denote a short description of the policy action indexed by i , and let $\mathcal{D}_i$ denote the collection of passages drawn from its contemporaneous policy documents. A text representation $\psi(\cdot)$ places the policy description and document passages in a common semantic space. The baseline retrieval rule retains the K passages closest to the policy description:

$$\mathcal{R}_K(q_i; \mathcal{D}_i) = \arg \min_{d \in \mathcal{D}_i}^K \text{dist}\{\psi(q_i), \psi(d)\}. \quad (6)$$

Retrieval settings. In the applications, I implement equation (6) using corpus-specific text representations. I use OpenAI’s *text-embedding-3-large* for the French VAT documents, OpenAI’s *text-embedding-3-small* for the reconstructed Cloyne sources, and Google’s *text-embedding-004* for the U.S. budget documents. The Romer–Romer validation starts from edited narratives and therefore does not require document retrieval. Documents are split into fixed-length chunks and the retrieval keeps the ten most relevant passages.

I use the French VAT corpus to assess retrieval quality in this fiscal-policy setting. Each VAT measure is manually linked to the Senate-report article discussing the corresponding budget-law pro-

³⁴An alternative would use the separate category probabilities as controls in a tax-policy reaction equation. That approach requires those probabilities and other relevant controls to capture systematic policy responses sufficiently well that the remaining tax variation is orthogonal to ε_{t+h} . Here I instead use the category probabilities to construct narrative support. Under calibration and within-type moment stability, variation in support recovers the exogenous-type outcome–tax moment without requiring the policy-response-type moment to be zero.

vision. Retrieval performance changes little across chunk lengths and retrieval depths. Among the ten highest-ranked passages, the correct article is retrieved for between 97 and 100 percent of measures across chunk lengths from 512 to 4,096 characters. With the baseline length of 2,048 characters, the correct article is retrieved in 97 percent of cases, with an average of 5.2 of the ten passages overlapping it.³⁵ I do not have an analogous hand-linked benchmark for the U.S. and UK corpora. I expect retrieval performance to be similar, although the appropriate settings may differ because annual U.S. budget documents and UK fiscal and parliamentary records are structured differently from French Senate reports.

Motivation extraction and representation. The retrieved passages are passed to the language models with a fixed prompt asking for the government’s stated motivations for the tax change. The prompt instructs the model to rely only on the retrieved excerpts, avoid adding interpretation, and return at most three motivations in a uniform bullet-point format. The dynamic fields in the prompt are the fiscal year, the tax type, and the policy description taken from the policy dataset.³⁶

Baseline implementation. For each policy action, the retrieval rule retains the ten passages most closely related to the policy description, and the baseline holds one extracted set of at most three motivations fixed. Let L_i denote the number of motivations retained for action i , and let $\hat{p}_{i\ell}^{(m,k)}$ denote the support assigned to motivation ℓ using representation model m and calibration sample k . The baseline policy-level score is

$$p_i = \frac{1}{L_i} \sum_{\ell=1}^{L_i} \left(\frac{1}{36} \sum_{m=1}^3 \sum_{k=1}^{12} \hat{p}_{i\ell}^{(m,k)} \right).$$

Each motivation is therefore evaluated under 36 representation-by-calibration mappings before the motivation-level scores are averaged within the policy action. The baseline averages across representation and calibration choices, but not across alternative motivation extractions.

Model variation and replication. The hidden-state representation stage uses three open-weight instruction-tuned models: *Gemma-4B*, *Mistral-7B*, and *Qwen-9B*. To assess whether the results depend on motivation extraction, I separately retain twenty fixed-seed stochastic motivation draws from each of Qwen-9B, Gemma-4B, and Mistral-7B. Holding retrieval fixed, I reconstruct policy-level support for each draw and re-estimate the macroeconomic responses. These alternative extractions are retained as an end-to-end sensitivity exercise rather than included in the baseline average. All retrieved passages, prompts, outputs, model identifiers, hidden-state files, and probability-mapping settings are stored so that each support score can be traced back to the source text, extracted motivations, calibration sample, and model calls. Appendix B.3 separates variation across draw-averaged extraction models from stochastic variation within a model.³⁷

³⁵Online Appendix OA.1.1.2 reports the complete comparison.

³⁶Online Appendix OA.1.1.3 reports the exact extraction instructions and the settings used in the open-weight extraction sensitivity exercise.

³⁷Online Appendix OA.1.2.4 reports the exact Hugging Face repositories, pinned revisions, and runtime environment used for replication.

3.3 Grounding and Placebo Diagnostics

Because the support score is generated from text, reproducibility and agreement across alternative constructions do not by themselves establish validity. Similar scores may reflect generic linguistic patterns or model priors rather than the passages associated with the specific policy action. Errors may enter through evidence retrieval, motivation extraction, or the subsequent probability construction. The diagnostics below examine whether the extracted motivations are grounded in the retrieved policy evidence.

The diagnostic builds on perturbation-based analyses of model behavior, which assess how outputs change when selected parts of the input are altered or withheld (Feng et al., 2018; Prabhakaran et al., 2019). The implementation is closest to Cohen-Wang et al. (2024): both vary which context sources are included and approximate the resulting changes in the model output with a linear surrogate. Here, the context sources are retrieved passages from policy documents, so the resulting contributions provide evidence linking an extracted motivation to specific parts of the documentary record.

Marginal passage contributions. For a policy action with K retrieved passages, let v_ℓ indicate whether passage ℓ is included when the extracted motivation is evaluated. I generate random passage-inclusion patterns $v = (v_1, \dots, v_K)$ while holding all other inputs fixed and use the linear approximation

$$\text{logodds}(v) \approx b + \sum_{\ell=1}^K w_\ell v_\ell,$$

where the left-hand side is the sequence-level log odds assigned to regenerating the motivation. The coefficients are estimated by OLS across the random inclusion patterns and summarize the contribution of each passage. A positive w_ℓ indicates that including passage ℓ increases the model’s support for the extracted motivation, while a coefficient near zero indicates limited dependence on that passage.

Placebo comparison. Positive passage contributions do not necessarily establish policy-specific grounding. For each policy action, I therefore repeat the exercise after adding semantically unrelated placebo passages drawn from the same document. I compare the contributions of the retrieved passages with the empirical distribution of contributions assigned to the placebo passages. If a motivation is grounded in the policy-specific evidence, the passages selected by retrieval should contribute more strongly than unrelated text from the same documentary source.

Table 2 reports the passage-level contributions and placebo comparisons for the 1977 French VAT reform.

I also apply a complementary event-level randomization diagnostic to the French VAT proposals.³⁸ These diagnostics assess whether the extracted motivations are anchored in the documentary record. They do not establish the macroeconomic identifying restrictions. They do not quantify how alternative support constructions affect the estimated macroeconomic responses; that sensitivity analysis is described in Subsection 4.3.

This construction provides action-level support scores p_i . Section 4 combines them with recorded revenue effects for quarterly macroeconomic estimation.

³⁸Online Appendix OA.1.1.5 formalizes this diagnostic and reports the results.

Table 2: Passage-Level Attribution Coefficients for the 1977 VAT Reform

	Retrieved segments (1–10)									
	1	2	3	4	5	6	7	8	9	10
M_1	41.56	20.92	7.33	-0.24	-1.76	8.36	2.78	-0.69	-4.93	4.47
M_2	18.82	-4.13	15.29	-4.56	0.25	1.20	1.93	-3.47	-0.51	-3.26
M_3	2.88	8.99	1.19	0.76	-1.26	1.04	1.05	-1.40	-0.46	2.67

Note: M1–M3 are the three motivations introduced in Section 2.2, in the order quoted. Each entry reports the OLS estimate of the corresponding passage’s marginal contribution to the sequence-level log odds assigned to regenerating the motivation. The coefficients are estimated from 64 random inclusion patterns over ten retrieved and thirty placebo passages. Shaded cells identify positive retrieved-passage coefficients that exceed every positive placebo coefficient for the same motivation. The shading is a descriptive coefficient-level comparison; Online Appendix OA.1.1.5 defines a complementary event-level randomization diagnostic.

4 Econometric Implementation

I combine the action-level support scores p_i with the source-reported revenue effects $\Delta\tau_i$. Here t indexes quarters: I first aggregate action-level information and then describe estimation and measurement sensitivity using the resulting quarterly series.

4.1 Policy-to-Quarter Aggregation

Let A_t denote the set of policy actions assigned to quarter t . Each action has a source-reported revenue effect $\Delta\tau_i$ and a policy-level support score $p_i \in [0, 1]$. The quarterly tax-policy measure is

$$\Delta\tau_t = \sum_{i \in A_t} \Delta\tau_i.$$

When at least one action is recorded, the quarterly support index is the absolute-tax-weighted average³⁹

$$p_t = \frac{\sum_{i \in A_t} |\Delta\tau_i| p_i}{\sum_{i \in A_t} |\Delta\tau_i|}.$$

Under the conditional calibration restriction stated in Appendix A.4.1, p_t is the expected share of gross tax variation in the quarter that belongs to exogenous policy actions. Equivalently, it is the probability that a randomly selected unit of gross tax variation comes from an exogenous action. It is not generally the probability that the entire quarter belongs to a single exogenous type. Nevertheless, $p_t = 1$ retains the interpretation that all policy variation in the quarter belongs to the exogenous type.

I retain quarters without a recorded policy action and set $\Delta\tau_t = p_t = 0$ in the estimation sample. Because no policy action is observed, the maintained model does not give p_t a latent-type interpretation in those quarters. Setting $p_t = 0$ is therefore a convention for completing the quarterly series, rather than a statement that no-action quarters belong to the policy-response type. This convention is nevertheless empirically consequential: it determines where no-action observations enter the affine moment relation and may therefore affect the variation used to recover the moments evaluated at $p = 1$. The fitted moment at $p = 0$ consequently reflects both no-action quarters and policy-action

³⁹Appendix A.4.1 shows when action-level calibration carries through this aggregation and states the corresponding within-type moment-stability conditions needed for identification with the quarterly support index.

quarters receiving little support for exogeneity.

Appendix F assesses the importance of this convention by assigning alternative common values between zero and one to no-action quarters and by estimating the relevant moments using recorded-action quarters only.

4.2 Support-Indexed Moment Estimation

This subsection gives the sample counterpart of the support-indexed moment in Section 1.4. For each horizon, I remove predictable variation in outcomes and tax changes using the controls in the empirical specification, estimate affine numerator and denominator moments, and evaluate both at $p = 1$. The application sections specify the samples, outcomes, horizons, and controls.

Estimation and inference. Fix a horizon h . Let $\mathcal{T}_h$ denote the usable quarterly sample at that horizon and let $D_t = \mathbb{1}\{|\Delta\tau_t| > 0\}$ indicate a quarter with nonzero net tax variation. Let $W_{t,h}$ collect a constant and the controls used in the application. I estimate the outcome and tax-change control projections over quarters in $\mathcal{T}_h$ for which $D_t = 1$:

$$\hat{\gamma}_{y,h} = \arg \min_{\gamma} \sum_{t \in \mathcal{T}_h} D_t (y_{t+h} - W'_{t,h} \gamma)^2, \quad \hat{\gamma}_{\tau,h} = \arg \min_{\gamma} \sum_{t \in \mathcal{T}_h} D_t (\Delta\tau_t - W'_{t,h} \gamma)^2.$$

The identifying restrictions are maintained for the population residuals corresponding to these projections, before the separate calendar-completion step described below. I then define zero-filled residuals

$$\tilde{y}_{t,h}^E = D_t (y_{t+h} - W'_{t,h} \hat{\gamma}_{y,h}), \quad \tilde{\Delta\tau}_{t,h}^E = D_t (\Delta\tau_t - W'_{t,h} \hat{\gamma}_{\tau,h}).$$

Thus the numerator and denominator observations

$$x_{t,h} = \tilde{y}_{t,h}^E \tilde{\Delta\tau}_{t,h}^E, \quad z_{t,h} = \left(\tilde{\Delta\tau}_{t,h}^E \right)^2$$

equal zero outside quarters with nonzero net tax variation. The affine projections themselves use every quarter in $\mathcal{T}_h$:

$$\hat{\theta}_{N,h} = \arg \min_{(a,b)} \sum_{t \in \mathcal{T}_h} (x_{t,h} - a - bp_t)^2, \quad \hat{\theta}_{D,h} = \arg \min_{(a_D,b_D)} \sum_{t \in \mathcal{T}_h} (z_{t,h} - a_D - b_D p_t)^2,$$

where $\theta_{N,h} = (a_h, b_h)'$ and $\theta_{D,h} = (a_{D,h}, b_{D,h})'$. No-action quarters therefore enter the second-stage projections as zero moment observations located at the baseline completion $p_t = 0$. This distinction between the residualization and affine-projection samples is maintained at every horizon. Below I suppress the superscript E on the zero-filled residuals for notational economy.

The estimated high-support response is

$$\hat{\beta}_h = \frac{\hat{a}_h + \hat{b}_h}{\hat{a}_{D,h} + \hat{b}_{D,h}}.$$

For the baseline U.S. application, inference uses a fully stacked Bartlett-HAC covariance for the control projections and the joint affine numerator and denominator, with bandwidth $L_h = \max\{4, h\}$. The

VAT application likewise includes country-specific residualization coefficients in its fully stacked covariance. The delta method is applied to each response ratio. The harmonized Romer–Romer comparison uses its paired fully stacked covariance, including the separate binary and probabilistic residualization coefficients. Appendix C records the application-specific implementation and provides the rank conditions, long-run covariance estimator, and response variance.

For compactness below, collect the first-step and affine parameters in $\eta_h = (\gamma'_{y,h}, \gamma'_{\tau,h}, \theta'_{N,h}, \theta'_{D,h})'$.

For later reference, let q_t denote any candidate support value and define the four-dimensional support-moment block

$$g_{t,h}^S(\eta_h; q_t) = \begin{pmatrix} (1, q_t)' [\tilde{y}_{t,h} \widetilde{\Delta\tau}_{t,h} - a_h - b_h q_t] \\ (1, q_t)' [\widetilde{\Delta\tau}_{t,h}^2 - a_{D,h} - b_{D,h} q_t] \end{pmatrix}, \quad (7)$$

where the residuals are evaluated at their estimated projection coefficients. For any complete support sequence $\mathbf{q} = (q_t)$ over the estimation sample, define

$$\bar{g}_h^S(\eta_h; \mathbf{q}) = \widehat{E}_h [g_{t,h}^S(\eta_h; q_t)].$$

The baseline sequence is $\mathbf{p} = (p_t)$; an alternative evaluator, calibration sample, or motivation extraction generates a complete alternative sequence $\mathbf{p}^{(r)} = (p_t^{(r)})$.

4.3 Sensitivity to Measurement Error in Narrative Support

The document-to-support mapping can be varied along three dimensions: the motivations extracted from the retrieved passages, the language model used to represent those motivations, and the synthetic sample used to calibrate the probability mapping. I first examine how these choices affect support scores and estimated responses. The baseline score averages across three representation models and twelve calibration samples. For the U.S. budget application, a joint exercise retains all $3 \times 20 \times 3 \times 12 = 2,160$ extraction-model, extraction-draw, representation-model, and calibration-sample combinations. These are alternative measurement constructions, not independent draws from a statistical model of measurement error.

Narrative support does not enter the first-step control projections. Changing p_t therefore leaves the residualized outcome and tax-change sequences fixed. In the U.S. budget specification, it changes the four affine support moments in $g_{t,h}^S$. The classical benchmarks and breakdown calculations below concern this application. The separate VAT construction-sensitivity exercise uses the pooled VAT specification and its own covariance procedure.

Corrected-moment benchmark. To ask what a classical correction would imply, write the measured quarterly score as $p_t = p_t^* + e_t$, where p_t^* is latent narrative support. This benchmark additionally assumes that e_t is conditionally mean zero and orthogonal to the variables entering the support-indexed moments. Let ω_t denote its conditional variance. The corrected support-moment block is

$$g_{t,h}^{S,ME} = g_{t,h}^S(\eta_h; p_t) + \omega_t(0, b_h, 0, b_{D,h})'. \quad (8)$$

The correction applies to the two moments in which support also enters as a weight (Evdokimov & Zeleneev, 2022). Its derivation and the additional measurement assumptions are given in Appendix C.7.

I first apply the correction to the baseline average of 36 constructions. I then extend the benchmark to the average of all 2,160 constructions, additionally treating extraction variation as classical measurement error. This extension changes both the measured support and the correction applied to its moments. To distinguish these effects, I report the response obtained with the expanded average before and after correction. The variance calculations allow dependence across constructions under the appendix’s stated restrictions. Fully stacked inference uses the corrected scores, including variation in the estimated correction statistic under the benchmark. Appendix D.2 reports both benchmarks alongside the construction-sensitivity results.

Approximate-moment sensitivity. To assess remaining distortions, including those shared across constructions, I allow the four identifying moments to deviate from zero and ask how large an adverse deviation would need to be to change the empirical conclusion. Let B_h be diagonal, with entries equal to the HAC standard deviations of the four support moments after accounting for the fitted control projections. The sensitivity set allows

$$\mathbb{E}[g_{t,h}^S(\eta_{h,0}; p_t)] = \frac{B_h \delta_h}{\sqrt{n_h}}, \quad \|\delta_h\|_2 \leq M. \quad (9)$$

The case $M = 0$ imposes exact validity of the baseline affine moments. Larger values permit greater misspecification relative to the sampling variability of those moments. Following Armstrong and Kolesár (2021), I construct pointwise confidence intervals allowing for the largest absolute local response bias within each radius.⁴⁰

I report separately the smallest radius at which the bias-aware 68% interval includes zero and the radius at which the point response can reach zero. The first is zero when the baseline interval already includes zero. For the response ratio, I distinguish the local linearized sign calculation from an exact calculation within the specified moment perturbation, and check the fitted denominator before interpreting a sign change.

The sensitivity radius describes how much the identifying moments are allowed to depart from zero, but its magnitude is easier to interpret against observed measurement changes. I therefore calculate the moment displacement induced by each alternative support construction and express it in the same standardized units. Holding the baseline parameter estimate fixed, define the moment displacement from construction r as

$$\Delta_h^{(r)} = \bar{g}_h^S(\hat{\eta}_h; \mathbf{p}^{(r)}) - \bar{g}_h^S(\hat{\eta}_h; \mathbf{p}).$$

Its equivalent sensitivity radius is

$$M_{h,\text{eq}}^{(r)} = \sqrt{n_h} \left\| B_h^{-1} \Delta_h^{(r)} \right\|_2. \quad (10)$$

Expressing observed measurement changes in these units provides a reference for interpreting the stress test, without treating those changes as bounds on shared error.

⁴⁰The baseline affine system is just identified and has no overidentification test. Even with additional moments, non-rejection need not exclude local misspecification that biases the response. Armstrong et al. (2025) study adaptation between restricted and more robust estimators; here I instead report sensitivity over posited bounds on moment departures.

The same displacement size can have different consequences depending on its direction. I measure its endpoint relevance by directional alignment: the absolute linearized response change divided by the largest absolute linearized change at the same radius. Alongside this local comparison, I report the response re-estimated under each construction. Appendix C.8 defines the calculation, and Appendix D.2 reports the results.

5 External Benchmarks and the Romer–Romer Comparison

This section first evaluates the proposed narrative-support measure against external benchmarks and then compares binary selection with probabilistic narrative identification in the Romer–Romer macroeconomic sample. The external exercises assess whether the measure preserves the ordering implicit in expert narrative classifications; because latent type is unobserved, they do not test the population calibration condition maintained in Section 1. Using edited Romer–Romer narratives, I ask whether the probability-construction stage assigns higher support to episodes classified as exogenous in the canonical U.S. series. Using Cloyne’s UK tax data, I ask whether the full document-grounded procedure aligns with an independent classification when retrieval, summarization, and motivation extraction are all performed from policy documents.

Expert narrative classifications inevitably involve difficult judgment calls, particularly for episodes with mixed or ambiguous motivations. Nevertheless, because the support measure is built from the same semantic distinction between categories that underlies the narrative literature, one would expect episodes classified as exogenous to receive stronger documentary support for exogeneity on average. Establishing this relationship is what makes the probability object informative for the budget-document application that follows.

5.1 Validation using the Romer–Romer narratives

This subsection applies the common procedure in Section 3 to the edited narratives used in Romer and Romer (2010). It first asks whether the procedure assigns systematically higher exogeneity support to RR-exogenous episodes. The subsequent macroeconomic comparison then asks how retaining probabilistic narrative information rather than applying the original binary selection rule changes the estimated response.

By conditioning on the Romer–Romer narratives, this benchmark abstracts from document retrieval while preserving the summarization, motivation decomposition, and probability aggregation stages of the framework. It therefore isolates the probability-construction stage before turning to the macroeconomic comparison.

5.1.1 Setup

The Romer–Romer replication material documents 50 major federal tax changes in the United States between 1945 and 2007. For each episode, the authors provide a detailed narrative discussion of the motivations underlying the reform, drawing on contemporaneous sources such as presidential speeches, budget documents, and Congressional records. Each narrative concludes with a classification of the tax change as either exogenous or endogenous.

To construct the benchmark validation task, I extract the narrative text describing each policy episode and remove the portion in which the authors explicitly state and justify their classification. In particular, I delete the sentence reporting the final classification along with the two immediately preceding sentences, which often contain direct cues revealing the intended label.⁴¹ The resulting edited narratives form the input to the validation exercise.

5.1.2 From edited narratives to policy-level probabilities

The edited RR narrative does not enter the benchmark as a single paragraph to be mapped directly into a binary label. Instead, each edited narrative is first summarized into a set of stated motivations. Because a single RR episode can contain several rationales, the same policy may generate multiple motivations, each of which is scored separately. These motivation-level outputs are then mapped into exogeneity probabilities and aggregated into a policy-level narrative-support probability.

The construction combines the probability mappings estimated across calibration sets and language models as described in Subsection 3.1. Throughout this benchmark, the object of interest is therefore the policy-level narrative support measure p_i , rather than the output from any single mapping.

5.1.3 Probability validation against RR classifications

Table 3 reports the original Romer–Romer binary label and asks whether RR-exogenous episodes receive systematically higher policy-level support. Mean support rises from 0.43 for RR-endogenous episodes to 0.60 for RR-exogenous episodes.

Table 3: Average Narrative Support by Original Romer–Romer Classification

Class	N	Mean p
Endogenous	24	0.43
Exogenous	26	0.60
Difference (exo - endo)	–	0.18

Note: Entries are computed from the 50 edited RR narratives. p is the baseline support probability averaged across twelve calibration samples and three representation models.

Mean support of 0.60 among RR-exogenous episodes remains well below one and, under the maintained calibration interpretation, implies material uncertainty about latent type.⁴²

5.1.4 Binary selection and probabilistic identification in the Romer–Romer sample

I next compare the traditional Romer–Romer binary design with the probabilistic framework. The binary benchmark retains only tax amounts attached to actions classified as exogenous by Romer and Romer. The probabilistic specification retains the broader set of RR tax actions and uses continuous narrative support to recover the response evaluated at $p = 1$. Both use the same outcomes, lag controls, timing, horizons, and HAC convention, but their control-fitting populations differ. The binary benchmark is controlled OLS on all usable quarters, with selected tax amounts set to zero

⁴¹Online Appendix OA.3.1.1 provides examples.

⁴²Online Appendix OA.3.4 relates intermediate probabilities to ambiguities in the original narratives, including current-macro motives, spending-package dependence, and deficit or social-insurance financing. Appendix D.1 reports probabilistic label-prediction diagnostics; Online Appendix OA.3.5 reports the training-leakage diagnostic.

outside retained actions; its regression residuals are not set to zero. The probabilistic specification fits the outcome and tax controls on recorded-action quarters and sets residual moment products to zero elsewhere. The tax series also differ by design because exclusion of RR-endogenous amounts is part of the binary identification procedure. This is a comparison of the two implemented designs, not a replication of the published RR distributed-lag estimator or an isolated test of the role of continuous support. For the probabilistic specification, the affine moment projections then use all usable calendar quarters, with no-action quarters assigned the baseline completion $p_t = 0$.

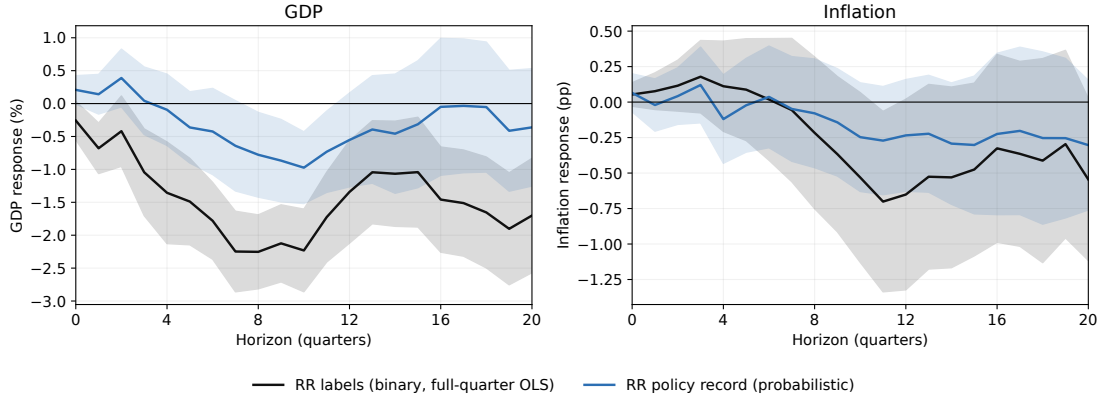


Figure 3: Traditional Romer–Romer Binary Selection and Probabilistic Narrative Identification

Note: The black line is controlled OLS on all usable quarters using the RR-label-selected tax series. The blue line evaluates the probabilistic response at $p = 1$, using recorded-action control fits and all-quarter affine moment projections with no-action support zero. Outcomes, lag controls, timing, and horizons are common; control-fitting populations and tax series differ. Each response uses its corresponding squared-tax normalization. Shaded areas are pointwise 68% confidence bands from the jointly stacked HAC covariance, including control estimation and numerator–denominator covariance.

Figure 3 shows a substantially smaller medium-run GDP contraction under the probabilistic framework. At horizon 8, the RR selected-shock response is -2.25 , compared with -0.78 for the continuous-support response. The latter is about 65 percent smaller in absolute value. For inflation, the estimated decline at horizon 8 is 0.22 percentage points under RR binary selection and 0.08 percentage points under probabilistic identification. Both responses are expressed per one-percentage-point-of-GDP tax increase. The paired 95 percent interval for the GDP difference, $[0.11, 2.84]$, excludes zero under this specification; the corresponding inflation interval includes zero. This inference conditions on the constructed support measure and does not establish that the difference is caused by correcting selection bias.

The smaller GDP response reflects its normalization rather than weaker negative GDP–tax comovement. Under the displayed moment scales, the probabilistic numerator is about 2.07 times as large in absolute value, but its squared-tax denominator is about 6.01 times as large. The larger normalization therefore outweighs the stronger negative comovement. Appendix D.1.2 reports the moment accounting. Because the averaging operators and control-fitting populations differ, these factors are numerical accounting, not a scale-invariant decomposition of causal bias.

The central medium-run attenuation is stable across the 22 alternative action-level constructions of narrative support: all produce a negative GDP response at horizon 8 that is smaller in absolute value

than the RR selected-shock estimate. The complete response path is less uniformly ordered, and the exact continuous endpoint is sensitive to how quarters without recorded actions enter the sparse RR support projection. Appendix D.1.2 documents the harmonized comparison, Appendix D.1.3 reports the construction diagnostic, and Appendix F reports the separate no-action completion exercise.

5.2 External Label Validation: Cloyne (2013) UK Tax Changes

The Romer–Romer exercise above isolates the probability-construction stage by starting from curated narrative text. I next ask whether the same probability object remains informative in a more demanding setting, where the relevant policy documents must first be reconstructed. Cloyne (2013) provides a useful benchmark for this purpose. The dataset records individual UK tax changes, their revenue effects and implementation dates, and an expert narrative classification of each change as endogenous or exogenous.

The validation starts from Cloyne’s row-level tax workbook. The reconstruction preserves the 2,462 tax-change rows in the original file and maps them into 85 narrative events. For each event, I build an event-level source base from contemporaneous UK fiscal and parliamentary materials, apply the retrieval and summarization procedure described above, and extract up to three stated motivations for each policy row. These motivations are then mapped into narrative-support probabilities. Because Cloyne’s classification is also defined at the policy-row level, it provides a direct external label against which to compare the resulting probabilities.

Table 4 asks whether rows classified as exogenous by Cloyne receive systematically higher support than rows classified as endogenous.

Table 4: Average Narrative Support by Cloyne Classification

Class	N	Mean p
Endogenous	795	0.72
Exogenous	1,306	0.82
Difference (exo - endo)	–	0.10

Note: Entries use 2,101 policy rows with usable Cloyne labels and complete probability inputs. No UK macroeconomic outcome enters the probability construction.

The Cloyne validation provides evidence about ranking, not calibration in levels or exact recovery of the binary labels. Cloyne-exogenous rows have average support of 0.82, compared with 0.72 for Cloyne-endogenous rows. Support is high in both classes, however, and the two distributions overlap substantially. Many Cloyne-endogenous rows receive high support because the reconstructed source text contains mixed package-level motivations, or because Cloyne’s row-level classification incorporates dependencies within a fiscal package that are not summarized by current-macro motives alone. The relevant result is the ordering across the two groups rather than the ability to reproduce Cloyne’s classification exactly. Appendix D.1.5 reports the corresponding one-variable label-prediction exercise.

Together, the two benchmarks show that the support measure preserves expert ordering when applied both to edited narratives and to reconstructed policy documents. The Romer–Romer comparison additionally shows that resolving that information through binary selection can materially affect the estimated macroeconomic response. I now use the resulting support measure to study tax transmission and the margins through which high-support tax changes operate.

6 Macroeconomic Effects of U.S. Budget Tax Proposals

I now apply the narrative identification framework to individual tax proposals reported in the annual *Budget of the United States Government*. Relative to existing narrative tax series, these documents provide substantially broader coverage of federal tax-policy proposals and retain the information needed to distinguish individual measures before they are consolidated into major historical episodes. They also contain the contemporaneous policy discussion needed to recover the administration’s stated motivations for each proposal. I use these data to estimate the macroeconomic response to proposed tax changes with high narrative support and to characterize the tax composition of the aggregate response.

6.1 U.S. tax data and narrative support

Existing narrative coverage. The narrative tax dataset of Romer and Romer (2010) identifies major federal tax changes in the United States over the postwar period. It covers large legislative episodes, including major tax reforms, Social Security legislation, and federal highway funding acts. The resulting series provides a well-defined measure of large, infrequent tax changes, but covers only a subset of federal tax-policy actions and records their composition at a relatively aggregated tax-instrument level.

Budget-based extension. I construct a complementary dataset from the annual *Budget of the United States Government* over 1950–2024. These documents report the administration’s proposed changes to federal taxation as part of the fiscal program submitted to Congress. For each fiscal year, I extract newly proposed tax measures at the highest level of detail reported in the Budget, retaining their estimated revenue effects and tax category whenever available. A proposed measure may introduce a new tax provision or modify an existing one, for example through an extension, acceleration, reduction, or expiration. The relevant observation is therefore the proposed policy change relative to prevailing tax policy, rather than whether the underlying tax provision itself is new.

The resulting data provide a substantially broader record of federal tax-policy proposals than the canonical narrative series and preserve the contemporaneous policy text needed to construct narrative support. The main application below uses the measures proposed in the annual budget documents.

The budget-based and Romer–Romer samples differ both in coverage and in the stage of the policy process at which tax changes are recorded. Romer and Romer record enacted tax changes, whereas the budget data record proposals made by the administration to Congress. The two sources can nevertheless overlap when a proposal reported in the Budget is subsequently enacted and enters the Romer–Romer series, or when a later budget proposes a modification of a provision associated with an earlier Romer–Romer episode. A single Romer–Romer episode may also correspond to several budget proposals distinguished by tax instrument or proposed effective date. Relating the two sources therefore requires a provision-level crosswalk rather than a mechanical match by law, date, or aggregate amount.⁴³

Table 5 summarizes the proposed tax measures extracted from the annual budgets. The six tax categories follow the classifications used in the source documents rather than an ex post taxonomy

⁴³Online Appendix OA.3.1.2 documents the clearest overlaps.

imposed by the researcher. For each measure, I record the proposed effective date reported in the Budget and assign the measure to the corresponding quarter. When no specific effective quarter is stated, I assign the measure to the first quarter of the fiscal year. The resulting dataset therefore describes changes in federal taxation proposed by the administration to Congress through the annual budget process, rather than the universe of enacted federal tax legislation.

Table 5: Distribution of Proposed Tax Measures by Category (United States)

Tax Category	# Proposed Measures
Individual Income Tax	371
Corporate Income Tax	361
Consumption Tax	238
Payroll / Social Insurance Tax	114
Other Receipts	89
Wealth Transfer Tax	17
Total	1190

Documentary evidence. The documentary corpus consists of the same annual budget volumes. In addition to reporting the administration’s proposed tax measures and their estimated revenue effects, the budgets discuss the fiscal program and the motivations underlying individual proposals. They therefore provide a contemporaneous documentary record from which to recover the stated rationale for the measures in the dataset. For each proposed measure, I use its label as the query against the corresponding budget volume and retrieve the passages most closely related to the proposal. The retrieved passages are then used to recover the stated motivations and construct narrative support following the common procedure in Section 3.

Quarterly variation in narrative support. I aggregate the proposed tax measures and their policy-level support scores to the quarterly frequency using the procedure in Subsection 4.1. This produces a quarterly budget-proposal series and a corresponding narrative-support index, p_t , which enter the support-indexed moment estimator.

The macroeconomic specification uses the 1947Q1–2007Q4 calendar and four lags each of log GDP, inflation, and the aggregate budget tax-policy measure. At each horizon, the outcome and tax measure are residualized over quarters with nonzero net tax variation, with residuals set to zero elsewhere; the affine numerator and denominator projections then use all usable calendar quarters. Appendix Table 13 records the complete specification.

Figure 4 reports the distribution of narrative support across quarters with budget-tax activity. The 1,078 matched measures correspond to 134 event quarters. Support is concentrated in the interior of the unit interval: its mean is 0.58 and its median is 0.59. The budget data therefore contain substantial variation in the strength of narrative support without naturally separating into fully exogenous and fully endogenous policy episodes.

Figure 5 shows the quarterly proposed revenue effects together with narrative support. The upper panel compares the budget-proposal series with the original Romer–Romer enacted exogenous tax series. The budget series contains 133 non-zero quarters, compared with 45 in the Romer–Romer

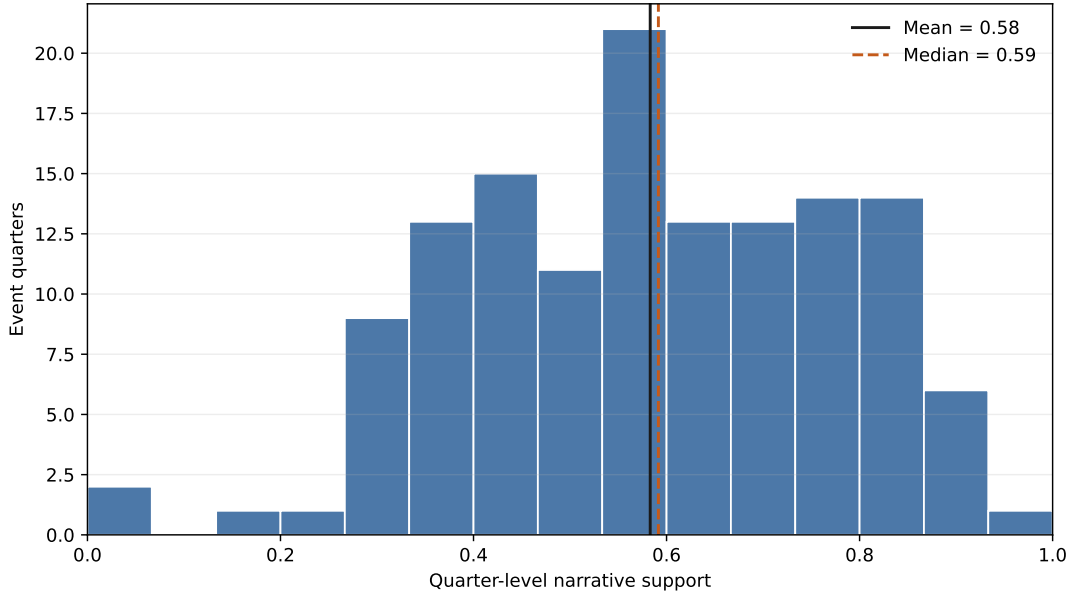


Figure 4: Distribution of Quarter-Level Narrative Support for U.S. Budget Tax Measures

Note: The figure uses U.S. budget measures with complete proposed effective dates, tax amounts, and narrative-support scores that can be matched to the quarterly macroeconomic sample. Policy-level support is aggregated to the quarter using absolute tax amounts as weights. The matched sample contains 134 quarters with budget-tax activity.

series, and the correlation between the two series is -0.03 . The lower panel reports p_t for the budget event quarters. Narrative support varies substantially over time and is nearly uncorrelated with the absolute size of the net quarterly proposed revenue effect, with a correlation of -0.04 .

6.2 Aggregate effects and tax composition

Figure 6 reports the estimated responses of GDP and inflation to a proposed tax increase at high narrative support, alongside the RR binary-selection benchmark. The budget estimates apply the support-indexed moment estimator in Section 1.4 to the granular budget-proposal record. At horizon 8, GDP is 1.87 percent lower and inflation is 0.57 percentage points lower. The GDP response reaches its largest decline of 2.24 percent at quarter 9, while the inflation response reaches its largest decline of 1.27 percentage points at quarter 11. Thus, across this broader policy record, high-support tax increases generate a sizable joint contraction in output and inflation.⁴⁴

The peak GDP contractions are close in the two applications, while the budget point estimates show a more pronounced medium-run decline in inflation. At horizon 8, GDP declines by 1.87 percent in the budget application and 2.25 percent under RR selection, while inflation declines by 0.57 and 0.22 percentage points, respectively. The RR benchmark uses the binary-classified tax series from Figure 3. This comparison describes responses across policy records that differ in coverage, tax composition

⁴⁴Appendix F varies the support assigned to no-action quarters over the unit interval and separately estimates the relevant moments using recorded-action quarters only. Across valid specifications, GDP remains negative at horizons 4, 8, and 12, and the medium-run decline in inflation is preserved. The large investment contraction reported below is also preserved.

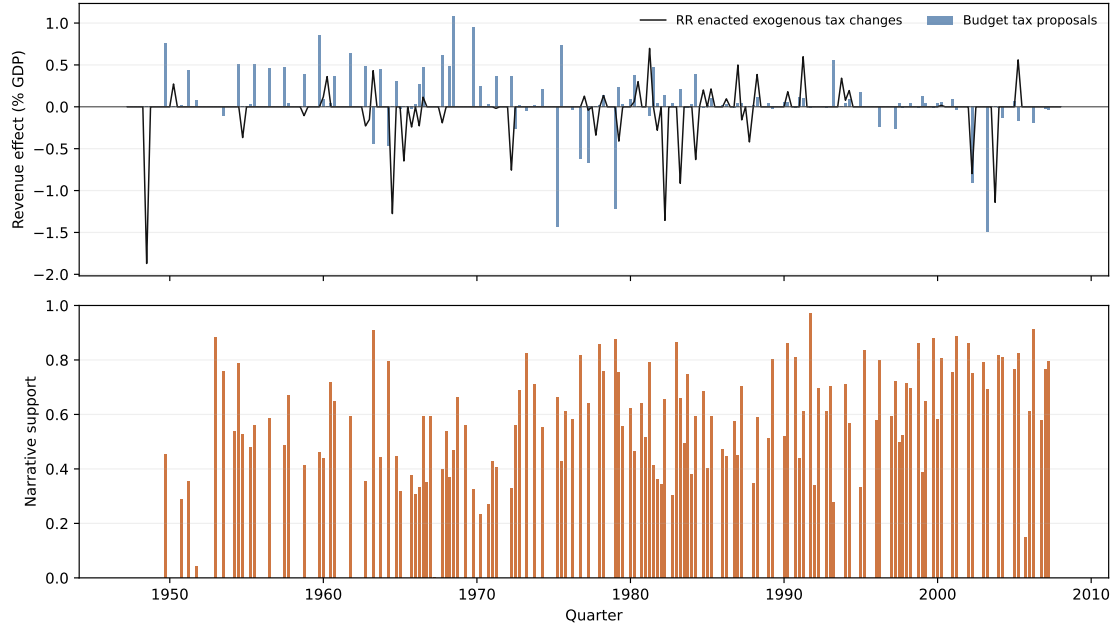


Figure 5: Quarterly Budget Tax Proposals and Narrative Support

Note: The upper panel plots proposed revenue effects for the budget series and enacted tax changes for the original Romer–Romer exogenous series, both expressed as a percentage of contemporaneous GDP. The lower panel reports the quarter-level narrative-support index p_t for quarters with budget-tax activity. “nzq” denotes the number of non-zero quarters.

and proposal-versus-enactment timing. I next examine which tax categories account for the aggregate budget contraction.

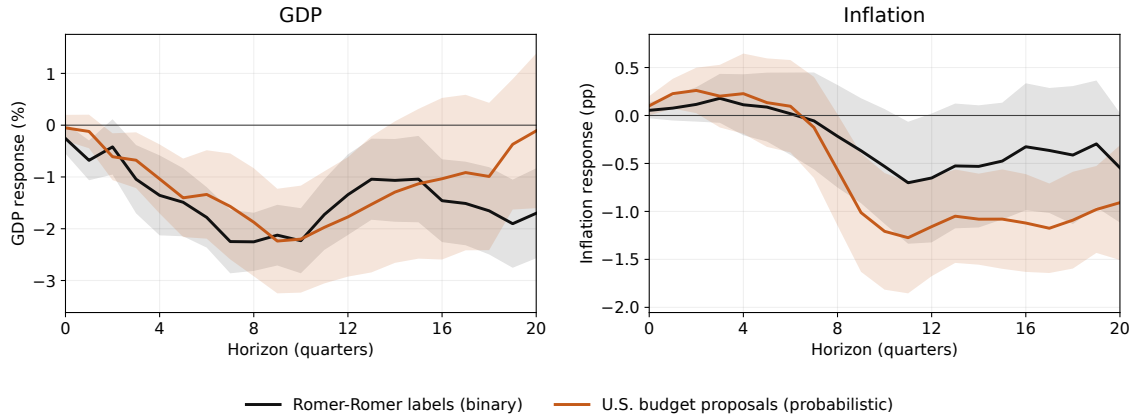


Figure 6: U.S. Budget Tax Proposals and the Romer–Romer Binary-Selection Benchmark

Note: Orange curves report the budget-proposal responses evaluated at $p = 1$, as in Section 1.4. Black curves report the Romer–Romer binary-selection benchmark from Figure 3. Both series are scaled to a tax increase of one percent of GDP. Shaded areas show pointwise 68% HAC confidence bands.

The aggregate GDP response combines tax proposals across several instruments. To assess which instruments account for the contraction, I decompose the aggregate support-indexed moment into

additive tax-category contributions. Let $\Delta\tau_{k,t}$ denote the quarterly proposed revenue effect associated with instrument k , so that

$$\Delta\tau_t = \sum_k \Delta\tau_{k,t}.$$

For each instrument, I estimate the outcome–tax moment using the aggregate narrative-support index p_t and normalize the resulting contribution by the same tax-variation moment used for the aggregate response. The contribution of instrument k at horizon h is

$$\hat{\beta}_{k,h}^{\text{add}} = \frac{\hat{a}_{k,h} + \hat{b}_{k,h}}{\hat{a}_{D,h}^{\text{all}} + \hat{b}_{D,h}^{\text{all}}}.$$

Because the decomposition uses a common support index and normalization, the contributions add to the aggregate response:

$$\hat{\beta}_h^{\text{all}}(1) = \sum_k \hat{\beta}_{k,h}^{\text{add}}.$$

These contributions characterize which tax instruments account for the estimated aggregate response in the historical sample; they do not separately identify causal responses to each instrument.

Table 6 reports the aggregate response and its tax-category contributions at horizons 4, 8, and 12. The medium-run GDP contraction is concentrated in individual and corporate income-tax proposals. At horizon 8, they contribute -2.01 and -0.86 percentage points, respectively, while the aggregate response is -1.87 : other categories partly offset the income-tax contraction. At horizon 12, their contributions remain negative, at -2.17 and -0.68 . Together, income-tax proposals account for 58.7 percent of gross recorded tax changes. The final column measures this share using absolute GDP-scaled projected revenue changes at the policy-action level, before quarterly netting; it describes the policy mix, not tax receipts or identifying weights. These contributions therefore reflect both the instruments represented in that mix and their outcome–tax moments. This exercise is distinct from Mertens and Ravn (2013), who separately identify personal and corporate income-tax shocks to estimate tax-specific causal responses.⁴⁵

The indirect-tax contributions are less coherent. Selective excises and bundles contribute a small decline, while energy, transport and other indirect taxes contribute positively at the main horizons. The first group includes familiar selective excises, generic bundles, and rate extensions; the second includes energy and producer-related taxes, transport measures, customs measures, and refunds. These proposals concern different tax bases rather than a single household consumption-tax wedge. Moreover, the record contains neither a broad federal VAT nor state and local sales-tax changes. The estimates therefore show which instruments carry the contraction in the observed federal budget mix; they do not show that consumption taxes are ineffective or that increasing them raises output. The separate UK–France VAT application provides a more narrowly defined consumption-tax contrast.

An additional joint-instrument exercise in Appendix D.3 estimates responses holding the other contemporaneous tax instruments fixed. Under stronger joint moment and exclusion restrictions, the income-tax GDP point estimates remain negative, although corporate-income estimates are imprecise.

⁴⁵Appendix D.4 compares the additive contributions and instrument-specific responses with binary selection at cutoffs 0.6, 0.7 and 0.8.

Table 6: Tax Composition of the U.S. GDP Response

Instrument	GDP contribution (percentage points)			Gross tax-change share (%)
	$h = 4$	$h = 8$	$h = 12$	
All-budget total	-1.04^* (0.67)	-1.87^{**} (1.05)	-1.77^* (1.16)	100.0
Individual income	-1.21^{**} (0.57)	-2.01^{**} (0.96)	-2.17^{**} (0.99)	36.4
Corporate income	-0.57^* (0.35)	-0.86^* (0.53)	-0.68^* (0.44)	22.3
Selective excises and bundles	-0.08 (0.15)	-0.24^* (0.19)	-0.12^* (0.10)	10.6
Energy, transport and other indirect taxes	0.70^{**} (0.34)	0.91^* (0.56)	1.10^{**} (0.56)	16.1
Other low-support taxes	0.12 (0.16)	0.32^* (0.25)	0.10 (0.22)	14.6

Note: Entries are additive contributions to the aggregate affine $p = 1$ endpoint for a proposed tax increase of one percent of GDP. All categories use the aggregate support index and common tax-variation denominator, so their contributions sum to the aggregate response up to rounding. Grey parentheses report fully stacked joint HAC standard errors, including fitted control projections and numerator–denominator covariance. * and ** indicate that the pointwise 68% and 90% confidence intervals exclude zero, respectively. The final column sums absolute GDP-scaled projected revenue changes at the policy-action level, before quarterly netting; these shares are not realized revenue shares or recovery weights. Entries are not separate instrument-specific multipliers.

6.3 Macroeconomic transmission

The aggregate contraction involves substantial adjustments in both consumption and investment. Figure 7 puts their responses in common GDP-contribution units, multiplying each estimated percentage response by its average nominal expenditure share in GDP. Investment falls substantially more than consumption in percentage terms, but its smaller expenditure share makes their GDP contributions comparable. At horizon 8, consumption contributes approximately -1.17 percentage points of GDP and investment -1.14 . Describing the contraction as investment-driven alone would therefore miss the role of consumption. This fixed-share accounting is a first-order comparison, not an exhaustive additive decomposition of GDP: other expenditure components and accounting differences remain. The pattern describes transmission of the composite historical policy mix, not an instrument-specific mechanism.

The income-tax-heavy composition of the aggregate endpoint provides useful context for the large proportional investment response. Earlier narrative evidence also finds that personal, corporate, and consumption taxes operate through different margins (Mertens & Ravn, 2013; Nguyen et al., 2021). Simulations from the calibrated model of Trabandt and Uhlig (2011) illustrate how different tax wedges can generate these contrasting expenditure patterns.⁴⁶ The U.S. estimates should be interpreted as the response to a composite of tax proposals, dominated at high narrative support by individual and corporate income-tax proposals, rather than as a common effect across tax instruments.⁴⁷

⁴⁶Online Appendix OA.5 reports the comparison and its normalization choices.

⁴⁷Online Appendix OA.4.2 reports additional expenditure components and comparisons with the Romer–Romer responses.

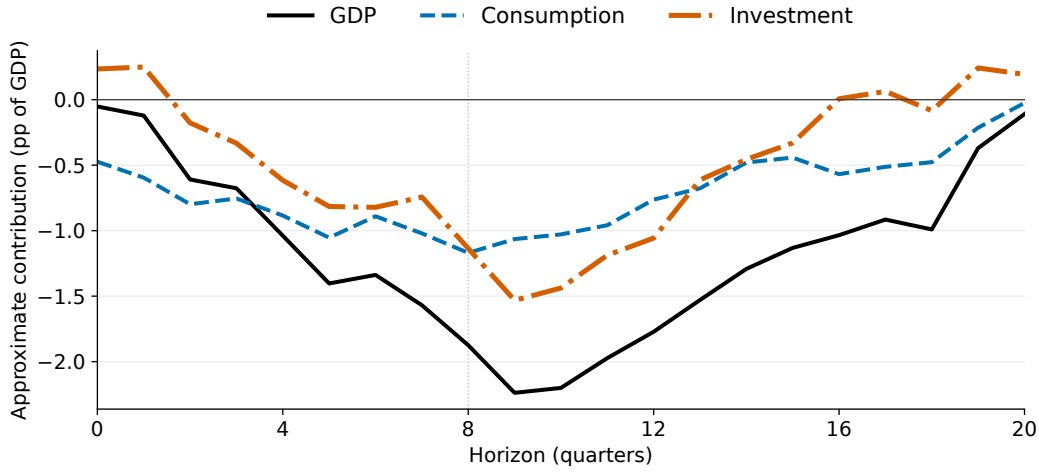


Figure 7: GDP and Domestic-Demand Contributions at High Narrative Support

Note: The figure reports the $p = 1$ responses underlying Figure 6. Consumption and investment log responses are multiplied by their sample-mean nominal expenditure shares, approximately 0.625 and 0.175, to express first-order contributions in percentage points of GDP. GDP is shown in percent. The figure displays point estimates; these expenditure contributions need not sum to the GDP response.

6.4 Sensitivity of the U.S. Estimates

Table 7 examines how the estimated responses change when narrative support is constructed differently. I vary the motivation-extraction model, extraction draw, representation model and calibration sample separately, and then vary all four components jointly. Among the individual components, GDP is most sensitive to the extraction model, while calibration-sample variation is relatively more important for inflation. Differences across representation models and repeated extraction draws are smaller.

Table 7: Sensitivity of U.S. Endpoint Estimates to Narrative-Support Construction

Construction varied	N	GDP		Inflation	
		Median	P90	Median	P90
Motivation-extraction model	3	0.256	0.320	0.060	0.087
Stochastic extraction draw	60	0.067	0.149	0.032	0.089
Representation model	3	0.114	0.129	0.024	0.040
Calibration sample	12	0.098	0.179	0.097	0.130
All stages jointly	2,160	0.267	0.605	0.144	0.326

Note: Entries report absolute changes in the re-estimated $p = 1$ response. The reference is the baseline construction except for stochastic extraction draws, which are compared with the corresponding extraction model's draw average. Each entry is the median across the horizon-specific statistics at horizons 4, 8 and 12; it is not a quantile pooled across horizons. GDP is measured in percent and inflation in percentage points. N is the number of constructions at each horizon. Rows vary one stage at a time except for the last row, which varies all stages jointly. These are measurement-sensitivity comparisons, not independent draws or confidence intervals.

Joint variation changes magnitudes but preserves the central medium-run pattern. Across horizons 4, 8 and 12, the median absolute change is 0.27 percentage points for GDP and 0.14 percentage points for inflation; the corresponding summaries of the 90th-percentile changes are 0.61 and 0.33 percentage points. GDP remains negative at all three horizons across all 2,160 constructions. Inflation is less

stable at horizon 4, where 75 percent of constructions retain the positive baseline sign, but is negative in all constructions at horizons 8 and 12.

I next ask whether interpreting these constructions under the classical repeated-measurement benchmark materially changes the estimated responses. The additional moment correction is small for both measurement averages. This finding is conditional on the classical restrictions, not evidence that total measurement error is small.

The measurement comparisons also help interpret the approximate-moment stress test. At horizons 4, 8 and 12, median GDP directional alignment ranges from 18 to 27 percent: observed construction changes typically have a much smaller absolute local effect than the most consequential distortion of the same standardized magnitude. This provides context for the stress test, without restricting the direction of errors shared across constructions. Appendix D.2 reports the confidence and point-sign breakdown thresholds.

7 The Indirect-Tax Contrast: VAT in the United Kingdom and France

The U.S. decomposition indicates that the contractionary aggregate contribution is concentrated in income-tax proposals. Although the budget record also contains substantial consumption and excise variation, these measures are heterogeneous and do not isolate a coherent indirect-tax experiment. I therefore turn to VAT policy changes in France and the United Kingdom as a more narrowly defined indirect-tax application.

VAT reforms are sparse, often proposed as part of broader tax packages, and frequently justified by several motives at once, including revenue needs, price stabilization, purchasing power, simplification, and competitiveness. These features make binary narrative methods especially restrictive. Applying the same support-indexed estimator to this specific tax base complements the broader U.S. budget application.

The VAT application asks whether document-based probabilities can turn this mixed narrative evidence into a macroeconomic object. A VAT increase raises the tax-inclusive price of consumption goods and lowers household purchasing power, making private consumption a natural margin through which the policy may affect domestic absorption. It may also generate a short-run inflation response and an external offset through lower imports. The empirical question is therefore not only whether VAT shocks reduce output, but whether the response has the mechanism one would expect from an indirect tax.

The United Kingdom and France provide a natural two-country setting. Both have long postwar VAT or VAT-equivalent histories and rich documentary records, while the United States does not have a federal VAT. The pooled application estimates the response to a high-support VAT increase using the same probability scale and GMM logic as the U.S. application.

7.1 VAT policy data and documentary sources

The French VAT data are built from the *Voies et Moyens Tome I*, the revenue annex to the annual draft budget law. This source decomposes the projected change in tax revenues into mechanical growth in the tax base, accounting-scope changes, the continuing effects of previous reforms, and newly proposed

measures. The application focuses on discretionary policy actions with direct macroeconomic content, rather than spontaneous revenue movements or scope transfers. In the full French corpus, the budget documents record 140 new VAT measures and 324 previous-measure entries. VAT is also quantitatively important: in the 2024 budget documents it accounts for about 38 percent of gross central-government tax revenues and about 6 percent of GDP.

The French proposals are assigned to their stated effective quarter or, when none is reported, to the first quarter of the budget year. The UK measures retain Cloyne’s implementation dates for announced budget measures. Source-reported revenue effects retain their sign and are scaled by the country-specific GDP normalization used in the panel application.

The documentary corpus for the French side consists primarily of contemporaneous Senate finance reports. These reports are useful because they discuss the budget articles in detail, record the government’s stated rationale, and often scrutinize the economic and fiscal consequences of the proposed measures. For the VAT corpus, each measure can be linked to the Senate-report discussion of the corresponding budget-law provision. This structure is used to audit retrieval quality in Subsection 3.2: the document retriever locates the relevant discussion for almost all VAT measures among the top retrieved passages. The extracted motivations are then mapped into the same robust support score used in the U.S. application.

For the United Kingdom, I use the action-level tax data underlying the Cloyne application and select VAT rows from tax-type and source labels containing VAT or value-added-tax language. The UK and French probabilities are placed on the same robust hidden-state calibration scale, so the panel uses a common support measure rather than country-specific binary classifications.

Table 8 summarizes the country-level inputs. The main VAT panel is the common UK–France sample from 1955Q1 to 2007Q4. It provides variation in recorded VAT actions and narrative support for estimating an instrument-specific response, while preserving the same quarter-level support construction across countries.

Table 8: VAT Policy Data Used in the UK–France Panel

Country	Main documentary source	Rows	nzq	Mean p	$p75$	$p90$
France	<i>Voies et Moyens</i> Tome I; Senate finance reports	122	39	0.71	0.81	0.88
United Kingdom	Cloyne action-level data; Budget, Hansard, Treasury record	168	72	0.80	0.91	0.94

Note: Rows count VAT policy actions and nzq counts nonzero VAT quarters in the common 1955Q1–2007Q4 sample. Support statistics are computed over nonzero VAT quarters after aggregating policy-level probabilities with absolute tax-change weights.

Figure 8 reports the resulting quarter-level support distribution. The object is not a country-specific binary label: event-quarter support is interior in both countries, with the UK distribution shifted toward higher support and the French distribution retaining more mass in the middle of the unit interval.

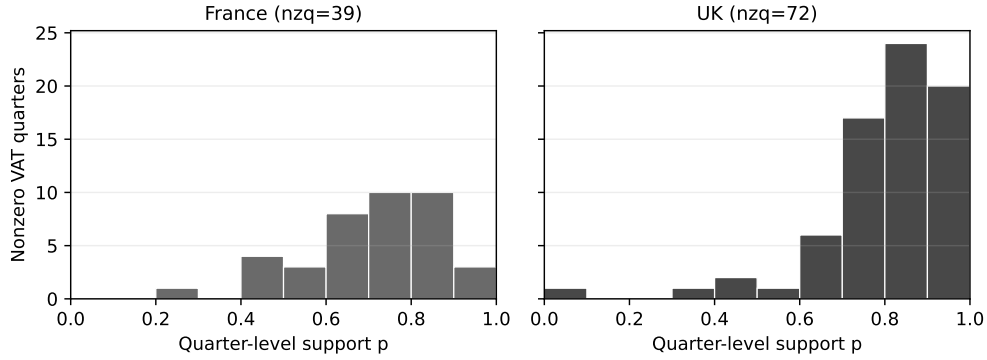


Figure 8: Quarter-Level Exogeneity Support for VAT Policy Actions

Note: The figure plots the distribution of quarter-level support p_{ct}^{VAT} for nonzero VAT quarters. Quarter-level support is the absolute-tax-weighted average of action-level support among VAT measures assigned to the same country-quarter.

The average support is high, especially in the United Kingdom. This is an empirical feature of the VAT narratives rather than a requirement of the method. Identification uses variation in narrative support and does not require many actions to receive probabilities close to one, or any action to be treated as certainly exogenous. Observations near one reduce the extrapolation required to evaluate the fitted moment at $p = 1$, but they do not substitute for calibration, support dispersion, and a well-behaved normalization. The high average is consistent with the nature of the VAT policy record. Many VAT changes are structural reforms of the tax system: introduction or coverage changes, rate unification, simplification, European harmonization, or long-run revenue design. They are therefore often documented as tax-design or fiscal-structure measures rather than as short-run stabilization responses. The probabilities should not be read as saying that every VAT action is an ideal isolated shock. They say that the stated motives often look closer to the exogenous/structural side of the narrative taxonomy than to contemporaneous macroeconomic management. This distinction also matters for Nguyen et al. (2021). The UK episodes they emphasize, including the VAT introduction transition in 1973Q2 and the 1979Q3 rate unification, are not assigned low support in this calibration: their quarter-level support is about 0.95 and 0.77, respectively. Their concern is therefore better viewed as a concern about anticipation, tax packages, and the isolation of the VAT component within broader consumption-tax measures, rather than as a prediction that these specific episodes should have near-zero narrative support.

Figure 9 translates the VAT policy rows into the quarterly policy series used in the estimation. The upper panels plot net quarterly VAT policy changes; the lower panels show the corresponding quarter-level exogeneity support. This is the direct UK–France analogue of the U.S. construction figure, but restricted to the VAT instrument.

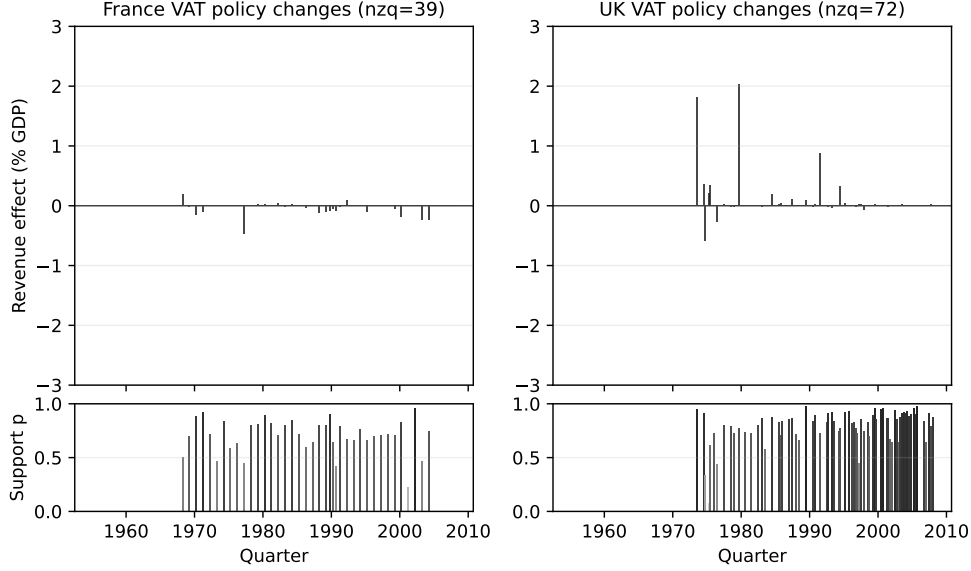


Figure 9: From VAT Measures to Quarterly VAT Policy Series in France and the United Kingdom

Note: The upper panels plot VAT policy rows aggregated to the country-quarter level, expressed as a percent of GDP. The lower panels report the corresponding absolute-tax-weighted quarter-level support. “nzq” denotes the number of nonzero VAT quarters in the common 1955Q1–2007Q4 sample.

7.2 Empirical design

The data are organized at the country-quarter level. In this subsection, $\Delta\tau_{ict}$ denotes the source-reported revenue effect for VAT policy row i in country c , quarter t , while $\Delta\tau_{ct}$ and p_{ct} denote the corresponding country-quarter VAT policy measure and support index. I aggregate them as

$$\Delta\tau_{ct} = \sum_{i \in (c,t)} \Delta\tau_{ict}, \quad p_{ct} = \frac{\sum_{i \in (c,t)} |\Delta\tau_{ict}| p_{ict}}{\sum_{i \in (c,t)} |\Delta\tau_{ict}|}, \quad (11)$$

I retain quarters without a recorded VAT action and set $\Delta\tau_{ct} = 0$ and $p_{ct} = 0$, following the baseline convention in Section 4.1. For quarters with recorded VAT actions, p_{ct} is the absolute-tax-weighted measure of narrative support; actions with larger revenue effects receive more weight.

The estimator is the pooled version of the high-support estimator used in the U.S. application. For each horizon, I residualize the outcome and VAT measure separately by country over quarters with nonzero net VAT variation. The GDP and expenditure specifications use four lags each of GDP relative to trend, inflation, and the aggregate all-tax measure; the inflation specification uses four lags each of GDP, inflation, and the aggregate all-tax measure. The residuals are set to zero elsewhere. I form $x_{ct,h} = \tilde{y}_{ct,h}^E \tilde{\Delta\tau}_{ct,h}^E$ and estimate over all usable country-quarters

$$x_{ct,h} = A_h + d_{c,h} (p_{ct} - 1) + u_{ct,h}. \quad (12)$$

The parameter A_h is the common numerator at $p = 1$. The country-specific slopes $d_{c,h}$ allow the support relation to differ across countries while imposing a common high-support numerator. For

the normalization, I estimate a separate affine squared-tax relation in each country, evaluate each at $p = 1$, and take their equal-weighted average. The response divides A_h by this pooled denominator and is interpreted as the response to a high-support VAT policy increase of one percent of GDP. Inference uses a fully stacked Bartlett-HAC covariance that jointly accounts for the country-specific residualization, numerator, and denominator moments, with calendar-quarter spacing and bandwidth $\max\{4, h\}$. Appendix Table 13 records the complete specification.

7.3 Macroeconomic mechanism

A VAT policy increase changes the tax-inclusive price of consumption goods and reduces household purchasing power. The relevant margins are therefore output, prices, private consumption, investment, and the external offset through imports and net exports. Figure 10 reports the main response. The top row shows output and prices; the bottom row shows the two domestic-absorption components that are most directly relevant for the VAT mechanism. A high-support VAT policy increase of one percent of GDP lowers GDP by about 2.5 percentage points of trend GDP after four quarters and by about 4.6 percentage points after eight quarters. Inflation rises initially and then returns toward zero, consistent with a VAT policy change that combines a direct tax-inclusive price effect with a contraction in real activity.

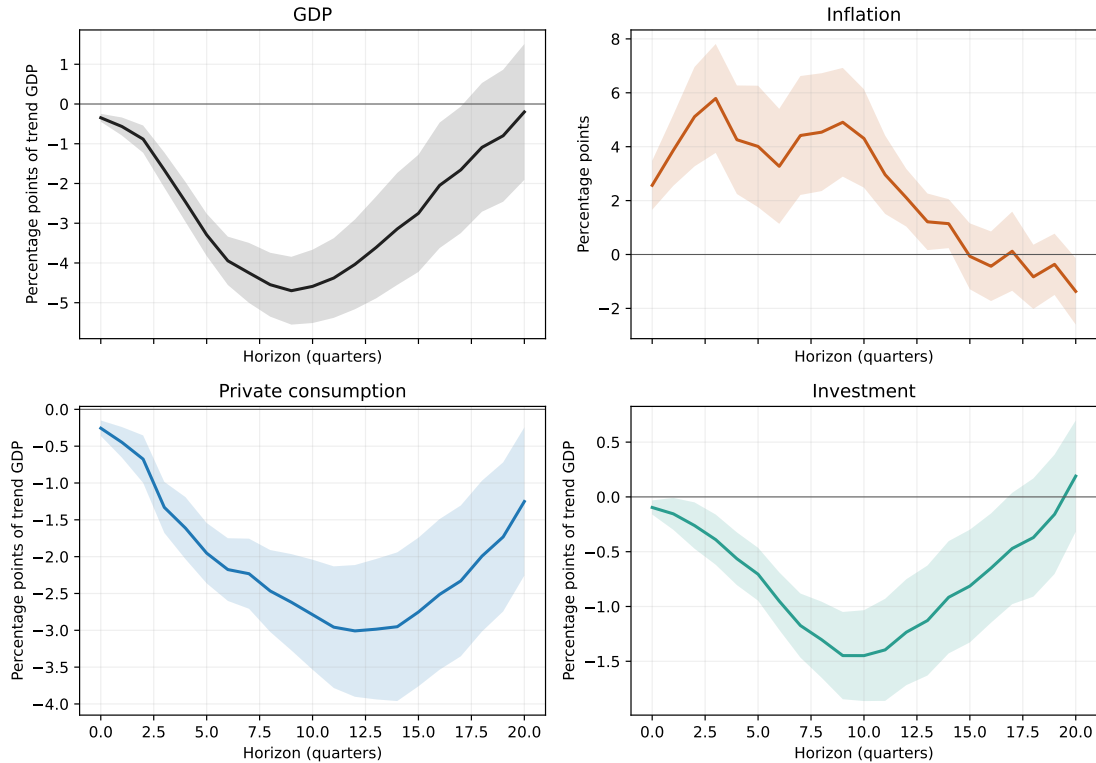


Figure 10: UK–France VAT: Output, Prices, and Domestic Absorption

Note: The figure reports the pooled high-support VAT endpoint. GDP, private consumption, and investment are expressed as $y_{c,t+h} - y_{c,t-1}$ in percentage points of trend GDP. Inflation is in percentage points. Responses are normalized by the equal-country average of the country-specific affine denominators evaluated at $p = 1$. Shaded areas are 68 percent fully stacked HAC confidence bands.

The response is concentrated in domestic absorption. Private consumption falls immediately and continues to decline through the medium run. Investment also falls, but its contribution in trend-GDP units is smaller than the consumption contribution. Table 9 summarizes this accounting at horizons 4, 8, and 12. At horizon 8, GDP is 4.6 percentage points below its pre-event trend-GDP position. The measured private-consumption contribution is 2.5 percentage points, while the investment contribution is 1.3. This contribution ranking partly reflects their different expenditure shares: in common units, consumption and investment fall by 6.6 and 11.2 log points, respectively. High-support VAT increases therefore raise prices initially and are followed by a broad contraction in domestic demand, with private consumption making the larger measured contribution to the output decline.

Table 9: VAT Main Mechanism at Key Horizons

Outcome	$h = 4$	$h = 8$	$h = 12$
GDP deviation from trend	-2.46	-4.55	-4.04
Inflation	4.26	4.54	2.11
Private consumption response	-1.61	-2.47	-3.01
Investment response	-0.56	-1.30	-1.23

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.

Simulations from the calibrated model of Trabandt and Uhlig (2011) illustrate how different tax wedges can generate these contrasting expenditure patterns.⁴⁸

The main VAT conclusions are also stable across alternative measurements of narrative support for recorded actions: at horizons 4, 8, and 12, all 22 constructions preserve the signs of the output and inflation responses and the private-consumption and investment contributions. Appendix E.1 reports the key-horizon estimates and links to the complete response paths in the online appendix.

Appendix F also varies the treatment of quarters without recorded VAT actions. The pooled recorded-action-quarter estimates and the alternative assignments $c = 0.5$ and $c = 1$ preserve the output contraction, decline in private consumption, and initial rise in inflation. Magnitudes are more sensitive than in the U.S. application, however, and the affine normalization becomes weak for some high- c assignments.

The additional expenditure components broaden the accounting interpretation.⁴⁹ Government consumption does not explain the initial contraction. Net exports move in the opposite direction and partly offset the fall in domestic demand, mostly because imports decline. The reported components are not an exhaustive additive decomposition of real GDP: inventories, statistical discrepancies, and chain-weighting differences remain in an accounting residual. The VAT application therefore complements the aggregate U.S. evidence with an output decline and an initial price response. Private consumption accounts for the largest domestic-demand contribution in trend-GDP units, even though the separate log-response estimates show that investment falls more proportionally.

A natural concern is that VAT changes may be bundled with other tax instruments. This is the central warning in Nguyen et al. (2021): a broad consumption-tax proxy can attribute package effects

⁴⁸Online Appendix OA.5 reports the comparison, its normalization choices, and its limitations.

⁴⁹Online Appendix OA.6.1 reports these components.

to indirect taxes if income-tax and consumption-tax changes are systematically linked. The same concern is relevant here, but the diagnostics point to a VAT-specific result. Same-quarter overlap is common because income-tax measures are frequent in budget quarters, yet the VAT-income correlation is small. Appendix E.2 shows that an explicit VAT/income adjustment leaves the GDP endpoint essentially unchanged and that dropping opposite-sign VAT-income package quarters does not weaken the response. By contrast, broadening the object from VAT to VAT plus duties and excises substantially attenuates the response. The result should therefore be read as a VAT-specific high-support endpoint, not as evidence about the broader average-consumption-tax-rate object studied by Nguyen et al. (2021).

8 Conclusion

This paper develops a probabilistic approach to narrative identification. Traditional narrative methods use documentary evidence to classify policy actions as exogenous or endogenous and retain a subset of observations for identification. I instead interpret the same evidence as probabilistic information about whether a policy action belongs to a latent exogenous type. Variation in narrative support for exogeneity then identifies the outcome-policy moment associated with that type. This formulation allows ambiguous and mixed-motive policy actions to remain informative without requiring any individual action to be treated as perfectly exogenous.

I implement this approach using contemporaneous policy documents and apply it to tax policy. The measurement procedure concentrates judgment in explicit choices over the narrative taxonomy, calibration design, and language models. The mapping is auditable, and the probability construction and estimation are reproducible conditional on the archived fixed motivations. The resulting probabilities are systematically related to the Romer–Romer and Cloyne classifications while retaining substantial variation within their binary categories. In the Romer–Romer sample, retaining this information rather than imposing binary selection yields a substantially more muted medium-run GDP point estimate. Applied at the level of individual U.S. federal tax proposals, the framework extends narrative identification to a substantially broader policy record and yields a sizable joint contraction in output and inflation. The aggregate endpoint is predominantly associated with individual and corporate income-tax proposals and is accompanied by a particularly large proportional investment decline. I then combine the UK tax record with a newly constructed narrative dataset of French VAT proposals from 1960 to 2024. The VAT application provides a complementary indirect-tax pattern: high-support VAT policy increases contract output and generate a short-run increase in inflation. Private consumption accounts for the larger domestic-demand contribution to the output decline, although investment falls more in percentage terms. Across these applications, probabilistic narrative evidence enters the identifying moments directly rather than first being resolved into a binary classification.

The same logic may be useful when the policy variation that is easiest to defend as exogenous is not representative of the variation of broader economic interest. Narrative studies of government spending, for example, often rely on military expenditures associated with geopolitical events, although their effects need not generalize to public investment or civilian purchases. Probabilistic narrative evidence could permit a broader set of spending actions to remain informative while allowing documentary support for exogeneity to vary across them. The approach could similarly use monetary-policy statements, minutes, and forecasts to distinguish different degrees of support for exogeneity across policy decisions.

An important extension concerns anticipation effects. The analysis characterizes why a policy action was taken but does not separately model when households and firms learned about it. A policy change may have strong narrative support for exogeneity while nevertheless being anticipated before implementation. Combining evidence about policy motivations with evidence about announcement and information timing would extend the framework to this additional identification problem.

References

- Armstrong, T. B., Kline, P., & Sun, L. (2025). Adapting to misspecification. *Econometrica*, 93(6), 1981–2005. <https://doi.org/10.3982/ECTA21991>
- Armstrong, T. B., & Kolesár, M. (2021). Sensitivity Analysis using Approximate Moment Condition Models. *Quantitative Economics*, 12(1), 77–108. <https://doi.org/10.3982/QE1609>
- Ash, E., & Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*, 15, 659–688. <https://doi.org/10.1146/annurev-economics-082222-074352>
- Bachmann, R., Born, B., Goldfayn-Frank, O., Kocharkov, G., Luetticke, R., & Weber, M. (2026). A temporary VAT cut as unconventional fiscal policy. *The Review of Economic Studies*, Article rdag055. <https://doi.org/10.1093/restud/rdag055>
- Barnichon, R., & Mesters, G. (2025, January). *Innovations Meet Narratives -Improving the Power-Credibility Trade-off in Macro* (BSE Working Paper No. 1475). Barcelona School of Economics. Retrieved November 27, 2025, from <https://bw.bse.eu/wp-content/uploads/2025/01/1475.pdf>
- Battaglia, L., Christensen, T., Hansen, S., & Sacher, S. (2024, December). Inference for Regression with Variables Generated by AI or Machine Learning. <https://doi.org/10.48550/arXiv.2402.15585>
- Beck, E., Eckert, F., Kuehne, L., Liebert, H., & Rosenblatt-Wisch, R. (2026). *Measuring economic outlook in the news* (Working Paper No. 2026-04). Swiss National Bank. https://www.snb.ch/en/publications/research/working-papers/2026/working-paper_2026_04
- Benzarti, Y., Carloni, D., Harju, J., & Kosonen, T. (2020). What Goes Up May Not Come Down: Asymmetric Incidence of Value-Added Taxes. *Journal of Political Economy*, 128(12), 4438–4474. <https://doi.org/10.1086/710558>
- Bini, L. (2025). *The macroeconomic effects of global supply chain shocks* [SSRN Working Paper]. <https://doi.org/10.2139/ssrn.5375776>
- Blanchard, O., & Perotti, R. (2002). An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output. *The Quarterly Journal of Economics*, 117(4), 1329–1368.
- Buckmann, M., & Hill, E. (2025, May). *Improving text classification: Logistic regression makes small LLMs strong and explainable ‘tens-of-shot’ classifiers* (Staff Working Paper No. 1127). Bank of England. Retrieved July 29, 2026, from <https://www.bankofengland.co.uk/working-paper/2025/improving-text-classification-logistic-regression-llms-tens-of-shot-classifiers>
- Carlson, J., & Dell, M. (2025, July). A Unifying Framework for Robust and Efficient Inference with Unstructured Data. <https://doi.org/10.48550/arXiv.2505.00282>
- Carrasco, M., & Florens, J.-P. (2000). Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 16(6), 797–834. <https://doi.org/10.1017/S0266466600166010>
- Cashin, D., & Unayama, T. (2016). Measuring Intertemporal Substitution in Consumption: Evidence from a VAT Increase in Japan. *The Review of Economics and Statistics*, 98(2), 285–297. https://doi.org/10.1162/REST_a_00531
- Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3), 305–334. [https://doi.org/10.1016/0304-4076\(87\)90015-7](https://doi.org/10.1016/0304-4076(87)90015-7)
- Chen, H., Didisheim, A., & Somoza, L. A. (2026, March). *Out of the Black Box: Uncertainty Quantification for LLMs via Conditional Probabilities* (NBER Working Paper No. w34965). National Bureau of Economic Research. <https://doi.org/10.3386/w34965>

- Clayton, C., Coppola, A., Maggiori, M., & Schreger, J. (2025, July). *Goeconomic Pressure* (NBER Working Paper No. 34020). National Bureau of Economic Research. <https://doi.org/10.3386/w34020>
- 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>
- Cohen-Wang, B., Shah, H., Georgiev, K., & Madry, A. (2024). ContextCite: Attributing Model Generation to Context. *Advances in Neural Information Processing Systems*, 37. <https://doi.org/10.52202/079017-3035>
- den Besten, T., Barnichon, R., Känzig, D. R., & Singh, A. (2026, April). *The macroeconomic effects of tariffs: Insights from 180 years of U.S. trade policy* (NBER Working Paper No. 35102). National Bureau of Economic Research. <https://doi.org/10.3386/w35102>
- Donald, S. G., Imbens, G. W., & Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics*, 152(1), 28–36. <https://doi.org/10.1016/j.jeconom.2008.10.013>
- Doronin, A., & Maiboroda, R. (2014). Testing hypotheses on moments by observations from a mixture with varying concentrations. *Modern Stochastics: Theory and Applications*, 1(2), 195–209.
- Evdokimov, K., & Zeleneev, A. (2022). *Simple estimation of semiparametric models with measurement errors* (tech. rep. No. CWP08/22). cemmap. <https://doi.org/10.47004/wp.cem.2022.0822>
- Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P., & Boyd-Graber, J. (2018). Pathologies of Neural Models Make Interpretations Difficult. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3719–3728). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1407>
- Friedman, M., & Schwartz, A. J. (1963). *A monetary history of the United States 1867-1960*. Princeton University Press.
- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, 57(3), 535–574. <https://doi.org/10.1257/jel.20181020>
- Hamilton, J. D. (1985). Historical Causes of Postwar Oil Shocks and Recessions. *The Energy Journal*, 6(1), 97–116.
- Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99(Supplement 1), S114–S133. <https://doi.org/10.1016/j.jinteco.2015.12.008>
- Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *The Quarterly Journal of Economics*, 133(2), 801–870. <https://doi.org/10.1093/qje/qjx045>
- Hartley, E. (2025, November). Narratives to Numbers: Large Language Models and Economic Policy Uncertainty. <https://doi.org/10.48550/arXiv.2511.17866>
- Hoberg, G., & Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, 124(5), 1423–1465. <https://doi.org/10.1086/688176>
- Jamilov, R. (2025). *Measuring Conflict Inflation* [Working paper]. <https://www.priceconflict.com/>

- Känzig, D. R., & Raghavan, R. (2026, July). *Supply chain shocks and the macroeconomy: Evidence from global shipping disruptions*. https://dkaenzig.github.io/diegokaenzig.com/Papers/kr_supplychainshocks.pdf
- Leeper, E. M., Walker, T. B., & Yang, S.-C. S. (2013). Fiscal foresight and information flows. *Econometrica*, 81(3), 1115–1145. <https://doi.org/10.3982/ECTA8337>
- Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, 30(1), 67–80. <https://doi.org/10.1080/07350015.2012.643126>
- Liubashenko, D., & Maiboroda, R. (2015). Linear regression by observations from mixture with varying concentrations. *Modern Stochastics: Theory and Applications*, 2(4), 343–353.
- Ludwig, J., Mullainathan, S., & Rambachan, A. (2026). Large Language Models: An Applied Econometric Framework. *Annual Review of Economics*, 18, 283–316. <https://doi.org/10.1146/annurev-economics-120925-105620>
- Maiboroda, R., Sugakova, O., & Doronin, A. (2013). Generalized estimating equations for mixtures with varying concentrations. *Canadian Journal of Statistics*, 41(2), 217–236.
- Martell, E. (2025). *Official Arm-Twisting? Measuring the Federal Reserve’s Use of Moral Suasion* [Job market paper]. <https://www.emilymartell.com/research>
- Mertens, K., & Ravn, M. O. (2012). Empirical Evidence on the Aggregate Effects of Anticipated and Unanticipated US Tax Policy Shocks. *American Economic Journal: Economic Policy*, 4(2), 145–181. <https://doi.org/10.1257/pol.4.2.145>
- Mertens, K., & Ravn, M. O. (2013). The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States. *American Economic Review*, 103(4), 1212–1247. <https://doi.org/10.1257/aer.103.4.1212>
- Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(4), 809–837.
- Nguyen, A. D., Onnis, L., & Rossi, R. (2021). The Macroeconomic Effects of Income and Consumption Tax Changes. *American Economic Journal: Economic Policy*, 13(2), 439–466. <https://doi.org/10.1257/pol.20170241>
- Prabhakaran, V., Hutchinson, B., & Mitchell, M. (2019). Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5740–5745). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1578>
- Ramey, V. A. (2011). Can Government Purchases Stimulate the Economy? *Journal of Economic Literature*, 49(3), 673–685. <https://doi.org/10.1257/jel.49.3.673>
- Ramey, V. A. (2016). Macroeconomic Shocks and Their Propagation. In *Handbook of Macroeconomics* (pp. 71–162, Vol. 2). Elsevier. <https://doi.org/10.1016/bs.hesmac.2016.03.003>
- Ramey, V. A., & Shapiro, M. D. (1998). Costly capital reallocation and the effects of government spending. *Carnegie-Rochester Conference Series on Public Policy*, 48, 145–194. [https://doi.org/10.1016/S0167-2231\(98\)00020-7](https://doi.org/10.1016/S0167-2231(98)00020-7)
- Ramey, V. A., & Zubairy, S. (2018). Government Spending Multipliers in Good Times and in Bad: Evidence from US Historical Data. *Journal of Political Economy*, 126(2), 850–901. <https://doi.org/10.1086/696277>

- Riera-Crichton, D., Vegh, C. A., & Vuletin, G. (2016). Tax multipliers: Pitfalls in measurement and identification. *Journal of Monetary Economics*, 79, 30–48. <https://doi.org/10.1016/j.jmoneco.2016.03.003>
- Rigobon, R. (2003). Identification through heteroskedasticity. *The Review of Economics and Statistics*, 85(4), 777–792. <https://doi.org/10.1162/003465303772815727>
- Rigobon, R., & Sack, B. (2004). The impact of monetary policy on asset prices. *Journal of Monetary Economics*, 51(8), 1553–1575. <https://doi.org/10.1016/j.jmoneco.2004.02.004>
- Romer & Romer. (1989). Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz. *NBER Macroeconomics Annual*, 4, 121–170. <https://doi.org/10.1086/654103>
- Romer & Romer. (2004). A New Measure of Monetary Shocks: Derivation and Implications. *American Economic Review*, 94(4), 1055–1084. <https://doi.org/10.1257/0002828042002651>
- 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>
- Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, 48(1), 1–48. <https://doi.org/10.2307/1912017>
- Stock, J. H., & Watson, M. W. (2018). Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments. *The Economic Journal*, 128(610), 917–948. <https://doi.org/10.1111/ecoj.12593>
- 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>
- Uhlig, H. (2005). What are the effects of monetary policy on output? Results from an agnostic identification procedure. *Journal of Monetary Economics*, 52(2), 381–419. <https://doi.org/10.1016/j.jmoneco.2004.05.007>
- Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., & Kong, L. (2022). Zerosen: Efficient zero-shot learning via dataset generation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11653–11669. <https://doi.org/10.18653/v1/2022.emnlp-main.801>

Appendix

A Identification Extensions and Policy Aggregation

This appendix examines within-type heterogeneity, state conditioning, the alternative share interpretation, quarterly aggregation, departures from calibration and stability, and the bias–variance comparison between binary selection and probabilistic recovery.

A.1 Within-Type Heterogeneity and Composition

The two latent policy types need not contain homogeneous actions. The following decomposition distinguishes changes in moments holding observed characteristics fixed from changes in their distribution within latent type.

Fix a horizon h and the population of policy observations used in the identifying argument. Let Z_t collect observed characteristics, such as motivation, magnitude, tax instrument, timing, or macroeconomic state. Let $F(\cdot | \cdot)$ denote the relevant conditional distribution of Z_t . All conditional statements below apply where the relevant conditioning values occur, and the displayed moments are assumed finite.

A.1.1 A General Decomposition

Iterated expectations give

$$\begin{aligned} & \mathbb{E}(y_{t+h}\Delta\tau_t | u_t = j, p_t = p) - \mathbb{E}(y_{t+h}\Delta\tau_t | u_t = j) \\ &= \int \left[\mathbb{E}(y_{t+h}\Delta\tau_t | u_t = j, p_t = p, Z_t = z) - \mathbb{E}(y_{t+h}\Delta\tau_t | u_t = j, Z_t = z) \right] dF(z | u_t = j, p_t = p) \\ & \quad + \int \mathbb{E}(y_{t+h}\Delta\tau_t | u_t = j, Z_t = z) \left[dF(z | u_t = j, p_t = p) - dF(z | u_t = j) \right]. \end{aligned} \quad (13)$$

The identity follows by adding and subtracting

$$\int \mathbb{E}(y_{t+h}\Delta\tau_t | u_t = j, Z_t = z) dF(z | u_t = j, p_t = p).$$

The first term in equation (13) captures changes associated with support after both latent type and the observed characteristics are held fixed. It vanishes if

$$\mathbb{E}(y_{t+h}\Delta\tau_t | u_t = j, p_t = p, Z_t = z) = \mathbb{E}(y_{t+h}\Delta\tau_t | u_t = j, Z_t = z). \quad (14)$$

The second term captures changes in the distribution of Z_t within latent type. It vanishes if that distribution is unchanged across support. An unchanged distribution is sufficient but not necessary: the distribution may change while leaving the weighted average moment unchanged.

Within-type moment stability requires the sum of the two terms in equation (13) to vanish. Conditions eliminating each term separately are therefore sufficient, not necessary. Changes holding characteristics fixed may average to zero or offset changes in composition. Conversely, equal average moments across values of Z_t eliminate the composition term but do not rule out support dependence holding Z_t fixed. Introducing Z_t for this decomposition does not impose calibration conditional on Z_t .

A.1.2 Policy-Response Motivations

The example in Section 1 is a discrete special case. Suppose $Z_t \in \{1, \dots, K\}$ records one of K motivations within the policy-response type. If equation (14) holds for $j = 0$, then

$$\begin{aligned} \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0, p_t = p) \\ = \sum_{k=1}^K \Pr(Z_t = k \mid u_t = 0, p_t = p) \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0, Z_t = k). \end{aligned} \quad (15)$$

Each probability is the fraction of policy-response actions with motivation k at support p . The motivation-specific moments may differ, including in sign. Stability requires their weighted average in equation (15), rather than the individual moments across motivations, to be the same at every support level.

For the two-motivation example, let $Z_t = 1$ denote demand management and $Z_t = 2$ spending-driven changes. Under equation (14),

$$\begin{aligned} & \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0, p_t = p) - \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0) \\ &= \left[\Pr(Z_t = 1 \mid u_t = 0, p_t = p) - \Pr(Z_t = 1 \mid u_t = 0) \right] \\ & \quad \times \left[\mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0, Z_t = 1) - \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0, Z_t = 2) \right]. \end{aligned} \quad (16)$$

Thus, when the two fixed motivation-specific moments differ, stability holds if and only if their relative frequency within the policy-response type is unchanged across support. With more than two motivations, changes in their relative frequencies can offset in the weighted average. Opposite signs alone neither invalidate stability nor ensure such an offset.

A.1.3 The Squared-Tax Moment and Orthogonality

Replacing $y_{t+h}\Delta\tau_t$ by $\Delta\tau_t^2$ in equation (13) gives the corresponding decomposition for the squared-tax moment. In particular, if squared-tax moments are stable across support after conditioning on motivation, then

$$\mathbb{E}(\Delta\tau_t^2 \mid u_t = 0, p_t = p) = \sum_{k=1}^K \Pr(Z_t = k \mid u_t = 0, p_t = p) \mathbb{E}(\Delta\tau_t^2 \mid u_t = 0, Z_t = k). \quad (17)$$

The benchmark requires this weighted average to remain unchanged as well. A change in composition that preserves the outcome-tax moment need not preserve the squared-tax moment. The same requirements apply within the exogenous type.

Under the common-response structural relation,

$$\begin{aligned} \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = j, p_t = p) &= \beta_h \mathbb{E}(\Delta\tau_t^2 \mid u_t = j, p_t = p) \\ & \quad + \mathbb{E}(\varepsilon_{t+h}\Delta\tau_t \mid u_t = j, p_t = p). \end{aligned} \quad (18)$$

This identity distinguishes support-related variation in policy magnitude from variation in policy-disturbance comovement. The two may offset in the numerator, so outcome-tax stability alone does not imply squared-tax stability. Joint stability of both moments makes the error-tax moment stable

within each type. Together with exogenous-type orthogonality, it therefore implies

$$\mathbb{E}(\varepsilon_{t+h}\Delta\tau_t \mid u_t = 1, p_t = p) = 0 \quad \text{for all } p \text{ in the support.}$$

This is a consequence of the joint moment restrictions and exogenous-type orthogonality, not of calibration. The corresponding policy-response-type error-tax moment may remain nonzero.

Observable comparisons. Differences in recorded motivations, instruments, timing, or other observed characteristics across support levels can reveal plausible sources of instability. Their distribution in the observed sample need not, however, coincide with their distribution within either latent type: observed composition can change simply because the relative incidence of the two types changes. Such comparisons therefore diagnose possible mechanisms of failure rather than establish the latent stability restriction.

A.2 Observed State Dependence

Outcome-tax moments may differ across macroeconomic states. For example, tax actions undertaken during periods of high unemployment may have different outcome-tax moments from actions undertaken during periods of low unemployment. If narrative support also differs across these states, the unconditional comparison may combine changes in latent-type composition with changes in economic conditions. When the relevant state is observed, the identification argument can instead be applied within states. The objective remains to recover the common causal response β_h .

Let Z_t now record an observed pre-policy macroeconomic state. Replace unconditional within-type moment stability with stability conditional on Z_t , as in equation (14), and its squared-tax counterpart. These restrictions allow the moments to differ across states but rule out further sorting by narrative support within each type and state. Additionally, impose conditional calibration:

$$\Pr(u_t = 1 \mid p_t = p, Z_t = z) = p.$$

This is stronger than baseline calibration. Iterated expectations then give

$$\begin{aligned} \mathbb{E}(y_{t+h}\Delta\tau_t \mid p_t = p, Z_t = z) &= p \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1, Z_t = z) \\ &\quad + (1 - p) \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0, Z_t = z). \end{aligned}$$

The same relation holds for the squared-tax moment. Provided sufficient variation in support remains within the relevant states, the two affine relations identify the exogenous-type moments at each state.

Averaging before normalization. Retain the common-response structural relation and the original orthogonality restriction $\mathbb{E}(\Delta\tau_t \varepsilon_{t+h} \mid u_t = 1) = 0$. This restriction need not hold separately in every state. To preserve its causal implication, average the recovered moments over the state distribution within the exogenous type before taking their ratio. Conditional calibration identifies this state distribution by assigning each observation weight $p_t / \mathbb{E}(p_t)$, provided $\mathbb{E}(p_t) > 0$. These weights recover the distribution of states within the exogenous type; the moments at each state are recovered through

the conditional identification argument above. Thus,

$$\beta_h = \frac{\int \mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1, Z_t = z) dF(z \mid u_t = 1)}{\int \mathbb{E}(\Delta\tau_t^2 \mid u_t = 1, Z_t = z) dF(z \mid u_t = 1)},$$

provided the denominator is positive. By iterated expectations, the numerator and denominator are the same exogenous-type average moments as in the baseline. The ratio at an arbitrary state need not equal β_h , since state-specific policy–disturbance comovement may be nonzero even when its exogenous-type average vanishes.

Centering in an affine state specification. The probability-weighted mean state is

$$\mathbb{E}(Z_t \mid u_t = 1) = \frac{\mathbb{E}(p_t Z_t)}{\mathbb{E}(p_t)}.$$

If the exogenous-type moments are affine in Z_t , recenter the state as $Z_t - \mathbb{E}(p_t Z_t)/\mathbb{E}(p_t)$. The exogenous-type endpoint of each moment, evaluated at centered state zero, then equals its average within the exogenous type. Their ratio therefore recovers β_h under the original orthogonality restriction. Centering at the unweighted mean state does not generally deliver these average moments.

With nonlinear state dependence, the recovered moments must instead be integrated over the exogenous-type state distribution. In either case, a ratio of average moments is not generally an average of state-specific ratios. The state distribution, calibration restriction, and orthogonality restriction must all refer to the same population of policy observations. In applications with controls, the moments are formed from the population residuals specified in the main text; residualization alone does not establish the conditional restrictions used here.

To interpret the state-specific ratios as causal responses, one would instead allow the structural response to vary with the state and impose exogenous-type orthogonality within each state, together with positive state-specific squared-tax moments.

A.3 Mixed Policy Actions and an Exogenous-Share Interpretation

The main text assigns each recorded tax-policy action to one of two latent types. Under a stronger alternative interpretation, a single recorded action may contain both an exogenous component and a policy-response component. Let $s_t \in [0, 1]$ denote the latent exogenous share relevant for the outcome–tax moment, so that these two components are represented by $s_t \Delta\tau_t$ and $(1 - s_t) \Delta\tau_t$. This proportional decomposition is an additional assumption; it is distinct from the aggregation of several binary policy actions considered below. The corresponding share calibration condition is

$$\mathbb{E}(s_t \mid p_t = p) = p.$$

Thus p is the average exogenous share among tax changes assigned support level p , rather than the probability that the entire tax change is exogenous.

Share calibration alone is insufficient for identification. The component associated with the exoge-

nous share must satisfy

$$\mathbb{E}[s_t \Delta \tau_t \varepsilon_{t+h} \mid p_t = p] = 0,$$

and the contaminated covariance per unit of remaining policy-response share must be stable:

$$\frac{\mathbb{E}[(1 - s_t) \Delta \tau_t \varepsilon_{t+h} \mid p_t = p]}{1 - p} \text{ does not depend on } p, \quad p < 1.$$

The first restriction defines the component indexed by s_t as exogenous. The second is the share-based analogue of within-type moment stability. Together, they imply that $\mathbb{E}(\Delta \tau_t \varepsilon_{t+h} \mid p_t = p)$ is proportional to $1 - p$ and therefore vanishes at $p = 1$. Combined with the analogous support-indexed restriction for the squared tax change maintained in Section 1, this gives the same response evaluated at $p = 1$ as the binary model.

This interpretation does not imply that $p_t \Delta \tau_t$ is an exogenous shock. Indeed,

$$p_t \Delta \tau_t = s_t \Delta \tau_t + (p_t - s_t) \Delta \tau_t,$$

so orthogonality of $s_t \Delta \tau_t$ does not imply orthogonality of $p_t \Delta \tau_t$. Calibration only requires $\mathbb{E}(s_t - p_t \mid p_t) = 0$; it does not require the discrepancy $s_t - p_t$ to be uncorrelated with $\Delta \tau_t \varepsilon_{t+h}$. The share interpretation therefore supports the moment-sorting argument, not an observation-level construction of the exogenous shock.

A.4 Quarterly Aggregation

A.4.1 Binary Policy Types and Quarterly Aggregation

The baseline model assigns each policy action a binary latent type. To distinguish the constituent policy actions from the quarterly aggregate, let $u_i = 1$ if action i is exogenous and $u_i = 0$ if it belongs to the policy-response type. For a quarter containing at least one policy action, define the realized exogenous share of gross tax variation as

$$\bar{u}_t = \frac{\sum_{i \in A_t} |\Delta \tau_i| u_i}{\sum_{i \in A_t} |\Delta \tau_i|}.$$

This definition does not make individual policy actions partially exogenous. Each u_i remains binary; $\bar{u}_t$ records only the composition of the actions assigned to the same quarter.

Marginal policy-level calibration, $\mathbb{E}(u_i \mid p_i) = p_i$, does not by itself imply calibration of the quarterly aggregate. A sufficient condition is that each policy-level score remains calibrated conditional on the tax amounts and the other scores entering the aggregation:

$$\mathbb{E}(u_i \mid \{p_j, \Delta \tau_j : j \in A_t\}) = p_i, \quad i \in A_t.$$

Under this condition,

$$\mathbb{E}(\bar{u}_t \mid \{p_i, \Delta \tau_i : i \in A_t\}) = \frac{\sum_{i \in A_t} |\Delta \tau_i| p_i}{\sum_{i \in A_t} |\Delta \tau_i|} = p_t.$$

It follows by iterated expectations that

$$\mathbb{E}(\bar{u}_t \mid p_t = p) = p.$$

Equivalently, consider selecting one unit of gross tax variation at random from the quarter, with each action selected in proportion to $|\Delta\tau_i|$. Conditional on the information used in aggregation, p_t is the probability that this unit belongs to an exogenous action. This is not generally the probability that every action in the quarter is exogenous. At $p_t = 1$, however, calibration and $\bar{u}_t \in [0, 1]$ imply $\bar{u}_t = 1$ almost surely, so all gross tax variation belongs to exogenous actions.

The calibration result concerns the composition of quarterly policy variation; it does not by itself imply the affine outcome–tax moment. That relation additionally requires that narrative support does not systematically sort the size, sign, composition, or timing of policy actions in ways that change the within-type outcome–tax and squared-tax moments. These are the within-type moment–stability and denominator restrictions maintained in Section 1. Under those restrictions, the $p = 1$ endpoint recovers the same exogenous-policy moment as in the binary benchmark. The gross share $\bar{u}_t$ need not coincide with the structural share s_t in Appendix A.3 without the additional proportional decomposition imposed there.

Finally, because $\bar{u}_t$ is constructed using absolute tax amounts while $\Delta\tau_t$ is the net quarterly tax-policy measure, $\bar{u}_t\Delta\tau_t$ need not equal the sum of exogenous action-level revenue effects when tax increases and cuts coexist. The same is true of $p_t\Delta\tau_t$. The quarterly support score is therefore used to sort moments and recover the $p = 1$ endpoint; it is not itself used to construct an observation-level exogenous tax shock.

A.5 Departures from Calibration and Within-Type Moment Stability

The affine benchmark combines two restrictions. Calibration determines the relative incidence of the exogenous and policy-response types at each support level, while within-type moment stability holds the relevant moments fixed within each type. This appendix considers departures from each restriction separately. Throughout, I retain the structural relation $y_{t+h} = \beta_h\Delta\tau_t + \varepsilon_{t+h}$ and the causal interpretation of the exogenous type.

A.5.1 Relaxing Calibration

Retain within-type moment stability, but allow the probability of the exogenous type to differ from the reported support score. Let

$$q(p) \equiv \Pr(u_t = 1 \mid p_t = p).$$

The baseline imposes $q(p) = p$. More generally, iterated expectations give

$$\mathbb{E}(y_{t+h}\Delta\tau_t \mid p_t = p) = q(p)\mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1) + [1 - q(p)]\mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0).$$

The observed moment is therefore affine in the true mixing probability $q(p)$, but need not be affine in the measured support score p . Nonlinearity in the support-indexed moment can thus arise even when the two type-specific moments remain stable.

If $q(\cdot)$ is known or identified from auxiliary information, the main identification argument carries through after replacing p_t by $q(p_t)$. In particular, weights satisfying

$$\mathbb{E}[w(p_t)q(p_t)] = 1, \quad \mathbb{E}[w(p_t)\{1 - q(p_t)\}] = 0$$

recover the exogenous-type moment. Under the analogous within-type stability condition for $\Delta\tau_t^2$,

applying the same construction to the squared tax change recovers the corresponding squared-tax moment and hence β_h .

If $q(\cdot)$ is unknown, variation in p_t alone does not generally determine the probability scale. A score may correctly rank actions by narrative support while mapping nonlinearly into the incidence of the exogenous type. The high-support endpoint also requires an appropriate normalization: interpreting $p = 1$ as the pure exogenous-type endpoint requires $q(1) = 1$, or an equivalent limiting condition when the endpoint is approached rather than observed. When observed support is bounded away from one, additional structure is needed to extrapolate the calibration map to that endpoint. Thus, relaxing calibration preserves the latent-component interpretation but shifts the identifying burden to the mapping from measured support into type probabilities.

A.5.2 Relaxing Within-Type Moment Stability

Now retain calibration, $\Pr(u_t = 1 \mid p_t = p) = p$, but allow the type-specific moments to vary with support. Then

$$\begin{aligned}\mathbb{E}(y_{t+h}\Delta\tau_t \mid p_t = p) &= p\mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 1, p_t = p) \\ &\quad + (1 - p)\mathbb{E}(y_{t+h}\Delta\tau_t \mid u_t = 0, p_t = p).\end{aligned}$$

Support now changes both the incidence of each type and the moments within types. One observed function of p therefore contains two unknown type-specific functions. Variation in support alone no longer separates the exogenous- and policy-response-type moments. In particular, an affine observed relation need not recover the exogenous-type moment once within-type stability is removed.

Separate recovery of the two exogenous-type moments is, however, stronger than necessary for identification of their causal ratio. A weaker sufficient restriction can be stated directly in terms of the policy–disturbance moment. Suppose

$$\mathbb{E}(\Delta\tau_t\varepsilon_{t+h} \mid p_t = p) = (1 - p)\kappa_h$$

for some constant κ_h . Any weight satisfying the recovery conditions in Section 1.3 then gives

$$\mathbb{E}[w(p_t)\Delta\tau_t\varepsilon_{t+h}] = \kappa_h \mathbb{E}[w(p_t)(1 - p_t)] = 0.$$

Using $y_{t+h} = \beta_h\Delta\tau_t + \varepsilon_{t+h}$ therefore yields

$$\beta_h = \frac{\mathbb{E}[w(p_t)y_{t+h}\Delta\tau_t]}{\mathbb{E}[w(p_t)\Delta\tau_t^2]},$$

provided the weighted squared-tax moment is nonzero. The numerator and denominator need not each equal their exogenous-type counterparts: support may sort tax magnitudes or other within-type features, provided the remaining policy–disturbance comovement has the stated form and is removed by the same weights.

This weaker restriction clarifies the role of the baseline assumptions. Within-type stability for $y_{t+h}\Delta\tau_t$ and $\Delta\tau_t^2$ is a transparent sufficient condition for recovering the two exogenous-type moments separately. It is not necessary when the target is only β_h and the relevant disturbance moment can instead be made orthogonal after weighting. Without either form of structure, however, relaxing

within-type stability leaves the component decomposition underidentified.

The two departures therefore have different implications. Relaxing calibration while retaining within-type stability preserves stable latent-component moments but changes how narrative support maps into their relative incidence. Relaxing within-type stability instead allows support to sort moments within each latent type and generally prevents component separation without additional restrictions. Nonlinear support-indexed specifications should therefore be interpreted according to which departure they represent and which restrictions preserve the causal endpoint or weighted ratio.

A.6 Binary Selection and Probabilistic Recovery: A Bias–Variance Comparison

This appendix compares binary selection with probabilistic recovery under the baseline restrictions, holding the policy series and control adjustment fixed and treating support as observed. The main Romer–Romer comparison instead changes the policy series as well.

Population discrepancy. The restrictions in Section 1.2 imply $\mathbb{E}(\Delta\tau_t\varepsilon_{t+h} \mid p_t = p) = (1 - p)\kappa_h$, where $\kappa_h = \mathbb{E}(\Delta\tau_t\varepsilon_{t+h} \mid u_t = 0)$; see also Appendix A.5. Consider a cutoff $\bar{p}$ that retains a positive fraction of observations with positive squared-tax moment. The selected-moment ratio differs from the causal response by

$$\begin{aligned}\mathcal{B}_h(\bar{p}) &\equiv \frac{\mathbb{E}(y_{t+h}\Delta\tau_t \mid p_t \geq \bar{p})}{\mathbb{E}(\Delta\tau_t^2 \mid p_t \geq \bar{p})} - \beta_h \\ &= \frac{\mathbb{E}(1 - p_t \mid p_t \geq \bar{p}) \kappa_h}{\mathbb{E}(\Delta\tau_t^2 \mid p_t \geq \bar{p})}.\end{aligned}\tag{19}$$

The discrepancy has the sign of κ_h and vanishes when $\kappa_h = 0$ or retained observations have support one. Under the full baseline, including squared-tax stability, higher cutoffs weakly reduce its absolute value. They retain fewer observations, but need not increase sampling variance: selection changes the informativeness of the retained variation as well as its quantity.

The weights in Section 1.3 instead cancel the policy-response contribution. Under the baseline restrictions and regularity conditions, probabilistic recovery consistently estimates β_h , but need not have smaller sampling variance.

Approximate sampling-risk comparison. Let n_h be the common underlying sample size and $V_h^B(\bar{p})$ and V_h^P the asymptotic variance factors for the fixed-cutoff and probabilistic estimators. The binary factor includes the retained fraction. An illustrative normal-approximation risk comparison, neglecting finite-sample ratio bias, is

$$\text{AMSE}_h^B(\bar{p}) \equiv \mathcal{B}_h(\bar{p})^2 + \frac{V_h^B(\bar{p})}{n_h}, \quad \text{AMSE}_h^P \equiv \frac{V_h^P}{n_h}.\tag{20}$$

Within this approximation, probabilistic recovery has lower risk when

$$\mathcal{B}_h(\bar{p})^2 > \frac{V_h^P - V_h^B(\bar{p})}{n_h}.\tag{21}$$

The comparison asks whether removing the population discrepancy compensates for any additional sampling variance. It requires well-behaved denominators and is not an exact finite-sample MSE ranking.⁵⁰

Minimizing the binary approximation over cutoffs characterizes an oracle comparison, not a rule implemented here. It requires identifying $\mathcal{B}_h(\bar{p})$, which binary selection alone does not provide, although the probabilistic restrictions can identify the necessary moments. The minimizing cutoff may differ across outcomes and horizons. Neither design has a universal finite-sample MSE advantage.

B Calibration Diagnostics

This appendix reports diagnostics for the synthetic calibration exercise used to construct the narrative-support probabilities. The external Romer–Romer and Cloyne exercises in Section 5 assess the economic content of the score; the diagnostics here assess the construction of the synthetic corpus and the stability of the resulting probability scale across calibration datasets and language-model representations.

B.1 Synthetic Holdout Performance

For each synthetic calibration set and each language model, I train the multinomial hidden-state decoder using five-fold stratified cross-validation within the synthetic corpus. The resulting diagnostics evaluate whether the category distinctions embedded in the synthetic taxonomy can be recovered from the hidden-state representations.

Table 10 reports average holdout performance across the twelve independently generated calibration sets. Performance is uniformly strong across models. Brier scores and log-loss values are low, classification accuracy is high, and the binary exogeneity classification implied by the taxonomy is almost perfectly separated in the synthetic environment.

Table 10: Synthetic Holdout Performance

Model	Brier	Log-loss	Accuracy	AUC	8-category log-loss
Gemma-4B	0.026	0.094	0.968	0.995	0.259
Mistral-7B	0.008	0.031	0.989	0.999	0.068
Qwen-9B	0.007	0.025	0.992	0.999	0.061

Note: Entries average five-fold stratified holdout performance across the twelve balanced synthetic calibration samples.

These results should be interpreted with caution. Synthetic examples are necessarily cleaner and more category-separated than real policy language. Strong synthetic performance therefore demonstrates that the category definitions are internally coherent and recoverable from hidden-state representations, but does not by itself establish external validity.

B.2 Stability Across Calibration Sets

The calibration design uses twelve separately generated datasets. Each dataset is used to train a separate decoder, producing twelve set-specific probability estimates for every policy action. Because

⁵⁰Any empirical variance comparison must include control-projection and joint numerator–denominator uncertainty using the appropriate stacked HAC systems in Appendix C. Uncertainty in measuring support is separate.

these estimates concern the same policy action and share the broader measurement architecture, they should not be interpreted as independent statistical replications. Their dispersion instead measures the sensitivity of the support score to the particular synthetic examples used during calibration.

Table 11 reports agreement across the twelve set-specific probabilities. Correlations are generally high, although meaningful variation remains in some samples. The median standard deviation across set-trained decoders ranges from roughly 0.08 to 0.11. The reported 90th percentile describes the upper tail of this dispersion. These statistics measure calibration-sample sensitivity, not distance from unobserved true support. The baseline averages the twelve scores.

Table 11: Agreement Across Calibration Samples

Sample	Mean Pearson	Mean abs. diff.	Median set s.d.	P90 set s.d.
U.S. RR replication	0.901	0.118	0.105	0.150
U.S. extension	0.788	0.133	0.108	0.186
France VAT	0.766	0.105	0.082	0.147
France non-VAT	0.747	0.095	0.076	0.136

Note: Pairwise agreement is computed across the twelve calibration-sample-specific support estimates after averaging over representation models. The last two columns summarize within-policy dispersion across samples.

B.3 Stability Across Models and Measurement Constructions

The baseline specification also averages probabilities across three open-weight language models: Gemma-4B, Mistral-7B, and Qwen-9B. This averaging is intended to reduce dependence on any particular hidden-state geometry.

Table 12 reports agreement across the model-specific probabilities after averaging over calibration samples. Pairwise correlations are generally high across applications, although the French samples exhibit somewhat greater dispersion than the U.S. samples. The lower panel separates variation across extraction models from stochastic extraction variation and reports the joint U.S. budget construction.

Table 12: Agreement Across Models and Measurement Constructions

Sample or source	N / variants	Mean Pearson	Mean abs. diff.	Median s.d.
U.S. RR replication	50	0.817	0.130	0.095
U.S. extension	1,188	0.806	0.129	0.091
France VAT	140	0.679	0.155	0.114
France non-VAT	1,020	0.753	0.138	0.103
Extraction model, draw-averaged	3	0.797	0.108	0.077
Stochastic draw within extraction model	20	0.926	0.050	0.042
All varied choices jointly	2,160	0.525	0.214	0.202

Note: The first four rows vary the language model used to represent a fixed motivation after averaging over calibration samples. The remaining rows use the U.S. budget application and vary motivation extraction. The joint construction combines extraction model, extraction draw, representation model, and calibration sample; its variants share components and are not independent draws.

The separate construction choices remain positively related, but their joint variation is materially larger than variation along any single dimension.

C Estimation and Inference for Support-Indexed Moments

This appendix records the sample construction and inference procedure for the high-support endpoint reported in Subsection 4.2. The baseline U.S. and VAT estimators use different samples at two stages: control projections are fitted over quarters with nonzero net tax variation, while the affine numerator and denominator projections use every usable calendar quarter. In quarters without a recorded tax action, I set both the tax change and narrative support equal to zero. Their zero-filled moment observations therefore enter the affine relation at $p_t = 0$; this is an estimation convention and does not define a separately observed policy class.

For baseline inference, the constructed support scores are treated as observed regressors. The sampling distribution describes uncertainty in the macroeconomic estimator conditional on the document-to-support mapping; it does not add uncertainty from reconstructing the support scores. The baseline U.S. bands, pooled VAT bands and paired Romer–Romer covariances include their application-specific outcome and tax residualization coefficients in fully stacked moment systems. The four affine moments conditional on those coefficients remain useful for displaying the estimator, but their covariance alone is not the reported baseline covariance. The classical benchmark below uses corrected scores under an additional measurement model.

C.1 Population Target and Sample Estimator

Fix a response horizon h , let $\mathcal{T}_h$ denote the ordered quarterly estimation sample, and let $n_h = |\mathcal{T}_h|$. Let $D_t = \mathbb{1}\{|\Delta\tau_t| > 0\}$, and let $W_{t,h}$ contain a constant and the application-specific controls. The first-stage estimates are

$$\hat{\gamma}_{y,h} = \arg \min_{\gamma} \sum_{t \in \mathcal{T}_h} D_t (y_{t+h} - \gamma' W_{t,h})^2, \quad \hat{\gamma}_{\tau,h} = \arg \min_{\gamma} \sum_{t \in \mathcal{T}_h} D_t (\Delta\tau_t - \gamma' W_{t,h})^2.$$

The implementation then forms the zero-filled residuals

$$\tilde{y}_{t,h}^E = D_t (y_{t+h} - \hat{\gamma}_{y,h}' W_{t,h}), \quad \tilde{\Delta\tau}_{t,h}^E = D_t (\Delta\tau_t - \hat{\gamma}_{\tau,h}' W_{t,h}),$$

and the numerator and denominator observations are

$$x_{t,h} = \tilde{y}_{t,h}^E \tilde{\Delta\tau}_{t,h}^E, \quad z_{t,h} = \left(\tilde{\Delta\tau}_{t,h}^E \right)^2.$$

For compactness, define

$$R_t = (1, p_t)', \quad \theta_{N,h} = (a_h, b_h)', \quad \theta_{D,h} = (a_{D,h}, b_{D,h})', \quad \iota = (1, 1)', \quad \vartheta_h = (\theta_{N,h}', \theta_{D,h}')'.$$

Provided that $\sum_{t \in \mathcal{T}_h} R_t R_t'$ is nonsingular,

$$\hat{\theta}_{N,h} = \left(\sum_{t \in \mathcal{T}_h} R_t R_t' \right)^{-1} \sum_{t \in \mathcal{T}_h} R_t x_{t,h}, \quad \hat{\theta}_{D,h} = \left(\sum_{t \in \mathcal{T}_h} R_t R_t' \right)^{-1} \sum_{t \in \mathcal{T}_h} R_t z_{t,h}. \quad (22)$$

Table 13: Empirical Implementation Record

Application	Calendar and horizons	Outcomes and tax-policy measure	Controls
Romer–Romer comparison	1947Q1–2007Q4; $h = 0, \dots, 20$; horizon-specific trimming for leads and lags	GDP in 100 log points and year-on-year GDP-deflator inflation in percentage points; RR tax amounts as a percent of GDP	Four lags each of log GDP, inflation, and the total RR tax-change series
U.S. budget application	1947Q1–2007Q4; $h = 0, \dots, 20$; horizon-specific trimming for leads and lags	Real quantity indexes in 100 log points and year-on-year GDP-deflator inflation in percentage points; proposed revenue effects as a percent of contemporaneous nominal GDP	Four lags each of log GDP, inflation, and the aggregate budget tax-policy measure
Pooled UK–France VAT	1955Q1–2007Q4; $h = 0, \dots, 20$; horizon-specific trimming within country	GDP and expenditure changes from $t - 1$, in percentage points of a quadratic trend in annualized real GDP; country-specific quarterly inflation series in percentage points; VAT changes as a percent of GDP	GDP and expenditure: four country-specific lags each of GDP relative to trend, inflation, and the aggregate all-tax measure; inflation: four lags each of GDP, inflation, and the aggregate all-tax measure
Application	Residualization sample	Affine projection and normalization	Inference
Romer–Romer comparison	Binary: all usable quarters; probabilistic: recorded-action quarters, with residual products zero-filled elsewhere	Binary selected-shock covariance ratio; probability affine projections use all usable quarters and are evaluated at $p = 1$; each response uses the denominator associated with its identifying tax series	Paired fully stacked Bartlett-HAC covariance, including separate binary and probabilistic residualization coefficients; bandwidth $\max\{4, h\}$
U.S. budget application	Quarters with nonzero net quarterly tax variation; residuals are zero-filled elsewhere	All usable quarters; affine numerator and squared-tax denominator evaluated at $p = 1$	Fully stacked Bartlett-HAC covariance, including outcome and tax residualization coefficients and joint numerator–denominator covariance; bandwidth $\max\{4, h\}$
Pooled UK–France VAT	Country-specific quarters with nonzero net VAT variation; residuals are zero-filled elsewhere	All usable country–quarters; common numerator endpoint with country-specific support slopes, divided by the equal-country average of country-specific $p = 1$ denominators	Fully stacked Bartlett-HAC covariance, including country-specific residualization coefficients and joint numerator–denominator covariance; bandwidth $\max\{4, h\}$

Note: “Usable” observations have finite outcomes, tax measures, support, and controls at the stated horizon. Thus effective start and end dates and the number of nonzero-net quarters vary with h . A recorded-action quarter can differ from a nonzero-net quarter when multiple actions within the quarter offset in the aggregate. French inflation is year-on-year GDP-deflator inflation; UK inflation is the quarterly series in the Cloyne replication macro data. Trend-GDP outcomes use rolling four-quarter real flows divided by a quadratic trend fitted to annualized real GDP.

The reported high-support response is

$$\hat{\beta}_h = \frac{\iota' \hat{\theta}_{N,h}}{\iota' \hat{\theta}_{D,h}} = \frac{\hat{a}_h + \hat{b}_h}{\hat{a}_{D,h} + \hat{b}_{D,h}}. \quad (23)$$

The four-dimensional affine score, conditional on the fitted control projections, is

$$g_{t,h}^A(\vartheta_h) = \begin{pmatrix} R_t(x_{t,h} - R_t' \theta_{N,h}) \\ R_t(z_{t,h} - R_t' \theta_{D,h}) \end{pmatrix}. \quad (24)$$

The affine estimator solves

$$\bar{g}_h^A(\hat{\vartheta}_h) = \frac{1}{n_h} \sum_{t \in \mathcal{T}_h} g_{t,h}^A(\hat{\vartheta}_h) = 0. \quad (25)$$

This conditional score describes the point estimator but does not, by itself, account for estimation of the control-projection coefficients. To state the U.S. inference, let $\eta_h = (\gamma'_{y,h}, \gamma'_{\tau,h}, \theta'_{N,h}, \theta'_{D,h})'$, and write $x_{t,h}(\gamma_y, \gamma_\tau)$ and $z_{t,h}(\gamma_\tau)$ for the zero-filled observations above evaluated at candidate control coefficients. The full score is

$$g_{t,h}^F(\eta_h) = \begin{pmatrix} D_t W_{t,h}(y_{t+h} - W'_{t,h} \gamma_{y,h}) \\ D_t W_{t,h}(\Delta \tau_t - W'_{t,h} \gamma_{\tau,h}) \\ R_t \{x_{t,h}(\gamma_{y,h}, \gamma_{\tau,h}) - R_t' \theta_{N,h}\} \\ R_t \{z_{t,h}(\gamma_{\tau,h}) - R_t' \theta_{D,h}\} \end{pmatrix}. \quad (26)$$

The sequential control projections and affine projections jointly solve $\bar{g}_h^F(\hat{\eta}_h) = n_h^{-1} \sum_{t \in \mathcal{T}_h} g_{t,h}^F(\hat{\eta}_h) = 0$. The first two score blocks are zero outside the nonzero-net-tax residualization sample. The last two blocks include every usable calendar quarter; their observed products are zero outside that sample, but their affine fitted values need not be. This construction retains the original quarterly spacing in the HAC calculation.

C.2 Identification of the Sample Moments

The two stages require distinct rank conditions. The event-sample control projections require $\Sigma_{W,h} = \text{plim } n_h^{-1} \sum_{t \in \mathcal{T}_h} D_t W_{t,h} W'_{t,h} \succ 0$. The affine coefficients require

$$\Sigma_{R,h} = \text{plim } \frac{1}{n_h} \sum_{t \in \mathcal{T}_h} R_t R_t' \succ 0.$$

Because $R_t = (1, p_t)'$, the latter condition is equivalent to a positive limiting variance of p_t over the affine-projection sample. It identifies both affine coefficient vectors in equation (22). Together, the two conditions make the full stacked Jacobian nonsingular: its diagonal blocks are $-\Sigma_{W,h}$ twice and $-\Sigma_{R,h}$ twice, with lower off-diagonal blocks reflecting the dependence of the affine moments on the control projections.

This rank condition concerns identification of the statistical affine projection. In the baseline implementation, quarters without a recorded policy action satisfy $p_t = 0$ by convention, so variation between event and non-event quarters may contribute to $\Sigma_{R,h}$. The structural interpretation of the

high-support endpoint additionally relies on variation in narrative support over policy variation, together with the calibration and within-type moment-stability restrictions in Section 1.4. In particular, the corresponding affine restriction must hold for both the numerator moment and the squared-tax denominator. Absent these restrictions, equation (23) remains a well-defined affine projection ratio evaluated at $p = 1$, but need not equal the structural response β_h . Appendix F assesses directly how the fitted response at $p = 1$ changes when the no-action completion is varied or when no-action quarters are excluded from the support-indexed projections.

C.3 Regularity Conditions

The asymptotic results are pointwise in the response horizon and conditional on the constructed narrative-support scores, but not on the estimated control projections. Thus h is held fixed while $n_h \rightarrow \infty$. Let

$$g_h^{F,0}(\eta_h) = \lim_{n_h \rightarrow \infty} E[\bar{g}_h^F(\eta_h)]$$

whenever this limit exists.

Assumption C.1 (Fixed-horizon regularity). For each fixed horizon h , the following conditions hold.

1. The parameter space is compact, $\eta_{h,0}$ lies in its interior, and $\eta_{h,0}$ is the unique zero of $g_h^{F,0}(\eta_h)$.
2. For some $\delta > 0$,

$$\sup_{t \in \mathcal{T}_h} E \left[\|g_{t,h}^F(\eta_{h,0})\|^{2+\delta} + \|\partial g_{t,h}^F(\eta_{h,0}) / \partial \eta_h'\|^{1+\delta} \right] < \infty,$$

and $p_t \in [0, 1]$. These bounds cover both the control-projection and affine scores.

3. A uniform law of large numbers holds for the full stacked moments and their first derivatives in a neighborhood of $\eta_{h,0}$. The sample Jacobian converges uniformly in probability to G_h^F .
4. The rank conditions $\Sigma_{W,h} \succ 0$ and $\Sigma_{R,h} \succ 0$ hold, so G_h^F is nonsingular.
5. The score sequence satisfies the central limit theorem

$$\frac{1}{\sqrt{n_h}} \sum_{t \in \mathcal{T}_h} g_{t,h}^F(\eta_{h,0}) \Rightarrow \mathcal{N}(0, \Omega_h), \quad (27)$$

where Ω_h is finite and positive semidefinite.

6. The population endpoint denominator is bounded away from zero:

$$\iota' \theta_{D,h,0} \geq \underline{q}_h > 0. \quad (28)$$

Assumption C.1 is stated at a high level to allow for changes in the unconditional distribution of the macroeconomic variables over the historical sample. It does not require strict stationarity. Standard weak-dependence conditions for time-series arrays, together with the displayed moment bound, provide primitive sufficient conditions for the uniform laws and central limit theorem. The substantive requirements for the fully stacked approximation are that the joint control-and-affine score sequence obeys a central limit theorem, that support retains nondegenerate variation, and that the endpoint denominator is separated from zero. Full rank of the event-sample control matrix is also required.

C.4 Asymptotic Distribution

Under Assumption C.1, standard just-identified GMM arguments imply that the full Jacobian has the block form

$$G_h^F = \begin{pmatrix} -\Sigma_{W,h} & 0 & 0 & 0 \\ 0 & -\Sigma_{W,h} & 0 & 0 \\ C_{Ny,h} & C_{N\tau,h} & -\Sigma_{R,h} & 0 \\ 0 & C_{D\tau,h} & 0 & -\Sigma_{R,h} \end{pmatrix},$$

where the cross-derivative blocks are the probability limits of the corresponding sample averages:

$$\begin{aligned} C_{Ny,h} &= -\text{plim} \frac{1}{n_h} \sum_{t \in \mathcal{T}_h} D_t R_t \widetilde{\Delta}_{t,h}^E W'_{t,h}, \\ C_{N\tau,h} &= -\text{plim} \frac{1}{n_h} \sum_{t \in \mathcal{T}_h} D_t R_t \widetilde{y}_{t,h}^E W'_{t,h}, \\ C_{D\tau,h} &= -2 \text{plim} \frac{1}{n_h} \sum_{t \in \mathcal{T}_h} D_t R_t \widetilde{\Delta}_{t,h}^E W'_{t,h}. \end{aligned}$$

These blocks carry first-stage estimation uncertainty into the endpoint covariance. At the population control coefficients, they need not vanish merely because the controls include a constant: the factors R_t vary with support. Consequently,

$$\sqrt{n_h} (\hat{\eta}_h - \eta_{h,0}) \Rightarrow \mathcal{N}(0, (G_h^F)^{-1} \Omega_h (G_h^F)^{-1'}) . \quad (29)$$

Define the response map

$$\beta_h(\eta_h) = \frac{\iota' \theta_{N,h}}{\iota' \theta_{D,h}}.$$

The estimator and target are obtained by evaluating this map at $\hat{\eta}_h$ and $\eta_{h,0}$, respectively. Then

$$\sqrt{n_h} [\hat{\beta}_h - \beta_h] \Rightarrow \mathcal{N}(0, \sigma_{\beta,h}^2), \quad (30)$$

where

$$\sigma_{\beta,h}^2 = \nabla \beta_h(\eta_{h,0})' (G_h^F)^{-1} \Omega_h (G_h^F)^{-1'} \nabla \beta_h(\eta_{h,0}). \quad (31)$$

Equation (29) follows from a mean-value expansion of the sample moments, the score central limit theorem, and Slutsky's theorem. The denominator condition (28) makes $\beta_h(\eta_h)$ continuously differentiable near $\eta_{h,0}$, so equations (30) and (31) follow from the delta method.

C.5 Long-Run Covariance Estimation

Let

$$\hat{g}_{t,h}^F = g_{t,h}^F(\hat{\eta}_h), \quad \hat{G}_h^F = \left. \frac{\partial \bar{g}_h^F(\eta)}{\partial \eta'} \right|_{\eta=\hat{\eta}_h}.$$

For $\ell \geq 0$, define

$$\hat{\Gamma}_{\ell,h} = \frac{1}{n_h} \sum_{t \in \mathcal{T}_h: t-\ell \in \mathcal{T}_h} \hat{g}_{t,h}^F \hat{g}_{t-\ell,h}^{F'}.$$

Here $t - \ell$ means ℓ calendar quarters earlier. Quarters excluded from the control projections have zero first-stage scores but retain their affine scores; excluded calendar gaps are not compressed when forming the HAC lags. The Bartlett-kernel HAC estimator is

$$\widehat{\Omega}_h = \widehat{\Gamma}_{0,h} + \sum_{\ell=1}^{L_{n,h}} \left(1 - \frac{\ell}{L_{n,h} + 1}\right) (\widehat{\Gamma}_{\ell,h} + \widehat{\Gamma}'_{\ell,h}). \quad (32)$$

For the asymptotic sequence, set

$$L_{n,h} = \max \left\{ 4, h, \left\lfloor n_h^{1/4} \right\rfloor \right\}. \quad (33)$$

For fixed h , this rule satisfies $L_{n,h} \rightarrow \infty$ and $L_{n,h}/n_h^{1/2} \rightarrow 0$.

Assumption C.2 (Long-run variance estimation). For each fixed h , the weak-dependence and moment conditions governing the score sequence are sufficient for the Bartlett estimator in equation (32), with bandwidth (33), to satisfy

$$\widehat{\Omega}_h \xrightarrow{p} \Omega_h. \quad (34)$$

Assumption C.2 is the standard HAC consistency requirement. It is satisfied, for example, under conventional mixing or near-epoch-dependence conditions with summable dependence and the moment bound in Assumption C.1. The bandwidth in (33) grows slowly enough to estimate increasingly distant covariance terms while remaining asymptotically negligible relative to the sample size. In the samples used in the paper, $\lfloor n_h^{1/4} \rfloor \leq 4$, so the rule coincides with the implemented choice $L_h = \max\{4, h\}$. Thus the asymptotic sequence does not alter the reported estimates or standard errors. Robustness checks vary the minimum bandwidth and its rate of increase.

The covariance estimator for the full parameter vector is

$$\widehat{V}_{\eta,h} = \frac{1}{n_h} (\widehat{G}_h^F)^{-1} \widehat{\Omega}_h (\widehat{G}_h^F)^{-1'}. \quad (35)$$

Under Assumptions C.1 and C.2, $\widehat{V}_{\eta,h}$ consistently estimates the pointwise asymptotic covariance of $\widehat{\eta}_h$, conditional on the constructed narrative support.

C.6 Inference for the High-Support Response

The gradient of the response map satisfies

$$\nabla_{\theta_N} \beta_h(\eta_h) = \frac{\iota}{\iota' \theta_{D,h}}, \quad \nabla_{\theta_D} \beta_h(\eta_h) = -\frac{\iota' \theta_{N,h}}{(\iota' \theta_{D,h})^2} \iota.$$

Equivalently,

$$\nabla \beta_h(\eta_h) = \begin{pmatrix} \mathbf{0}_{\dim(\gamma_{y,h})} \\ \mathbf{0}_{\dim(\gamma_{\tau,h})} \\ \iota / (\iota' \theta_{D,h}) \\ -(\iota' \theta_{N,h}) \iota / (\iota' \theta_{D,h})^2 \end{pmatrix}. \quad (36)$$

The estimated variance of the response is

$$\widehat{V}_{\beta,h} = \nabla \beta_h(\widehat{\eta}_h)' \widehat{V}_{\eta,h} \nabla \beta_h(\widehat{\eta}_h). \quad (37)$$

The resulting standard error accounts jointly for the estimated control projections, the affine numerator and denominator, and their covariance. The four-moment covariance based on equation (24) conditions on the estimated residuals; the fully stacked covariance also incorporates uncertainty from their estimation.

The same variance can be written using the estimated influence score

$$\widehat{\psi}_{t,h} = -\nabla \beta_h(\widehat{\eta}_h)' (\widehat{G}_h^F)^{-1} \widehat{g}_{t,h}^F. \quad (38)$$

Applying equation (32) to $\{\widehat{\psi}_{t,h}\}_{t \in \mathcal{T}_h}$ yields the scalar long-run variance entering (37).

The delta-method approximation requires the denominator evaluated at $p = 1$ to be well separated from zero. Accordingly, the estimation diagnostics report $\widehat{a}_{D,h} + \widehat{b}_{D,h}$, its standard error, and the fitted denominator over the empirical support of p_t . Delta-method bands should not be interpreted at horizons for which this condition is empirically doubtful; such cases require inference designed for a weakly identified ratio.

C.7 Corrected Moments for Generated Support

To distinguish the generated score from latent narrative support in this subsection, write

$$p_t = p_t^* + e_t, \quad E(e_t \mid p_t^*, x_{t,h}, z_{t,h}) = 0, \quad \omega_t = E(e_t^2 \mid p_t^*, x_{t,h}, z_{t,h}).$$

The classical-error benchmark also requires the residual measurement error to be orthogonal to the structural residuals in the numerator and denominator projections. Let

$$\vartheta_h = (a_h, b_h, a_{D,h}, b_{D,h})'.$$

Conditional on the residualization coefficients, the four corrected affine moments are

$$g_{t,h}^{\text{ME}}(\vartheta_h) = \begin{pmatrix} x_{t,h} - a_h - b_h p_t \\ (x_{t,h} - a_h - b_h p_t) p_t + b_h \omega_t \\ z_{t,h} - a_{D,h} - b_{D,h} p_t \\ (z_{t,h} - a_{D,h} - b_{D,h} p_t) p_t + b_{D,h} \omega_t \end{pmatrix}, \quad E[g_{t,h}^{\text{ME}}(\vartheta_{h,0})] = 0. \quad (39)$$

This is the componentwise form of the compact correction in equation (8). The first and third entries are the intercept moments; the second and fourth are the support-weighted moments affected by measurement error in p_t .

The correction follows directly from the quadratic expansion used by Evdokimov and Zeleneev (2022). For example, the second numerator moment has the form

$$g_x(p) = (x_{t,h} - a_h - b_h p) p, \quad g_x''(p) = -2b_h.$$

The first-order term generated by e_t has mean zero. The second-order term shifts the naive moment by $-b_h\omega_t$, which is removed by the correction in equation (39). Because $g_x(p)$ is quadratic, the expansion is exact under the maintained additive classical-error assumptions. The same argument applies to the denominator support moment.

Exactness here concerns the population moment correction under the stated assumptions, not finite-sample unbiasedness of the estimated response ratio.

Baseline measurement average. Let $p_i^{(m,k)}$ denote the score for action i from representation model $m \in \{1, 2, 3\}$ and calibration sample $k \in \{1, \dots, 12\}$. Each construction is first aggregated using the baseline absolute-tax weights:

$$p_t^{(m,k)} = \frac{\sum_{i \in A_t} |\Delta\tau_i| p_i^{(m,k)}}{\sum_{i \in A_t} |\Delta\tau_i|}, \quad p_t = \frac{1}{36} \sum_{m=1}^3 \sum_{k=1}^{12} p_t^{(m,k)}.$$

The variance calculation therefore applies to the quarterly objects used in estimation, without assuming independent measurement errors across actions in the same quarter. Under the benchmark, write

$$p_t^{(m,k)} = \mu_t + \xi_{t,m} + \zeta_{t,k} + r_{t,mk},$$

with mutually uncorrelated mean-zero model, calibration and residual effects, and a common variance within each component. The untruncated estimate of the variance of the average is

$$\begin{aligned} \hat{\omega}_t &= \frac{\hat{\sigma}_{\text{model},t}^2}{3} + \frac{\hat{\sigma}_{\text{calibration},t}^2}{12} + \frac{\hat{\sigma}_{\text{residual},t}^2}{36} \\ &= \frac{MS_{\text{model},t} + MS_{\text{calibration},t} - MS_{\text{residual},t}}{36}. \end{aligned} \tag{40}$$

The MS terms are the ordinary mean squares in the balanced two-way decomposition. This combination is unbiased under the stated error structure. Component estimates, and potentially the combined estimate, can be negative because they involve differences of mean squares. Setting negative components to zero introduces upward bias; I retain that truncation only as a diagnostic. In the U.S. benchmark, the combined estimates are nonnegative.⁵¹ The baseline zero-support convention is retained in quarters with zero net tax variation, including quarters without recorded actions; their correction statistic is also zero.

Expanded measurement average. The expanded benchmark additionally varies extraction model and stochastic draw. I first aggregate every construction to quarters, then average the twenty draws within each extraction model. Write $p_t^{(e,m,k)}$ for these draw averages, where e indexes the three extraction models. Their grand mean uses all 2,160 original constructions. The benchmark assumes that these cells share the same latent mean and that their measurement errors are uncorrelated when all three indices differ. Errors in cells sharing an index may be correlated. The classical orthogonality conditions on the controlled outcome-tax and squared-tax variables are also maintained in this benchmark.

Let $\mathcal{D}$ contain the ordered pairs with $e \neq e'$, $m \neq m'$ and $k \neq k'$, and let $\bar{p}_t$ be the grand mean of

⁵¹Truncating components changes corrected responses by at most 0.000042 across horizons 0–20.

the draw averages. The local variance statistic is

$$\hat{\omega}_t = \bar{p}_t^2 - \frac{1}{|\mathcal{D}|} \sum_{((e,m,k),(e',m',k')) \in \mathcal{D}} p_t^{(e,m,k)} p_t^{(e',m',k')}.$$

Under the disjoint-cell restriction, the expected product in the second term is the square of latent support, whereas the expected squared grand mean also contains the variance of its measurement error. Their difference therefore estimates that variance. Averaging draws within extraction models respects their nesting; it does not treat matching draw numbers across models as a common crossed factor or the 2,160 cells as independent measurements.

The disjoint-cell restriction is an assumption, not a property established by the model labels. Shared architectures, calibration designs or documentary omissions can generate dependence even when indices differ. A distortion common to every construction is not identified by this variance calculation.

Inference for the corrected benchmark. The estimated $\hat{\omega}_t$ enters the quarterly corrected scores together with the observed support and controlled variables. I stack these scores with the control-projection normal equations and use the corresponding Jacobian, calendar-spaced HAC covariance and response-ratio gradient. This includes the local correction statistic’s variation and covariance with the other scores under the stipulated classical and time-series assumptions. It is neither held fixed in uncorrected scores nor added as an independent variance term. This calculation does not integrate over an unknown distribution of shared training or calibration decisions. I check corrected support variation, rank and denominator positivity before reporting a ratio.

C.8 Approximate-Moment Sensitivity

The approximate-moment exercise is complementary to the classical correction. In the reported application it is centered on the baseline four-dimensional affine moment block in equation (7) and allows those moments to deviate locally from zero:

$$E[g_{t,h}^S(\eta_{h,0}; p_t)] = \frac{B_h \delta_h}{\sqrt{n_h}}, \quad \|\delta_h\|_2 \leq M. \quad (41)$$

Here, B_h is diagonal, with entries equal to the HAC standard deviations of the four moment contributions after profiling out the estimated control projections. Thus each coordinate of δ_h describes a possible violation in units of the sampling variability of the corresponding moment. The vector δ_h determines how the violation is distributed across the numerator and denominator moments, while M bounds its joint magnitude. The case $M = 0$ imposes exact moment validity. For $M > 0$, the confidence interval allows for the largest absolute *linearized* endpoint bias over the permitted moment distortions, following Armstrong and Kolesár (2021). The baseline sampling variance and the moment scales are held fixed as M varies.

The sensitivity radius can also be related to a concrete perturbation of the support scores. Let Δ_h denote the resulting displacement of the four population moments. Its equivalent standardized magnitude is

$$M_h^{\text{eq}} = \sqrt{n_h} \|B_h^{-1} \Delta_h\|_2. \quad (42)$$

For classical error with average variance $\bar{\omega}$, the leading displacement of the two support moments is

proportional to $-b_h\bar{\omega}$ and $-b_{D,h}\bar{\omega}$. More generally, equation (42) can translate the moment changes induced by an alternative language model, calibration set, or motivation extraction into the M scale used in the sensitivity figures.

The equivalent radius measures the norm of the complete standardized displacement, not its effect on the endpoint. To make this distinction explicit, let

$$G_h^S = E \left[\frac{\partial g_{t,h}^S(\eta_{h,0}; p_t)}{\partial \vartheta'_h} \right]$$

denote the Jacobian of the four support moments with respect to $\vartheta_h = (\theta'_{N,h}, \theta'_{D,h})'$, conditional on the residualization coefficients, and let $\beta_h(\vartheta_h)$ denote the response ratio evaluated at $p = 1$. Define

$$\kappa_h = -(G_h^S)^{-'} \nabla \beta_h(\vartheta_{h,0}).$$

For a small observed displacement Δ_h , the resulting first-order endpoint change is $\kappa'_h \Delta_h$. Among all displacements with equivalent radius M_h^{eq} , the largest first-order endpoint change is

$$\frac{M_h^{\text{eq}}}{\sqrt{n_h}} \|B'_h \kappa_h\|_2.$$

The directional alignment reported in the application is therefore

$$\rho_h^{\text{align}} = \frac{|\kappa'_h \Delta_h|}{M_h^{\text{eq}} \|B'_h \kappa_h\|_2 / \sqrt{n_h}}. \quad (43)$$

Directional alignment lies between zero and one. A value near one means that the observed displacement has a large absolute local effect relative to other displacements of the same standardized size; a value near zero means that its local effect is small. This is an absolute-change comparison, not specifically a measure of movement toward zero. Re-estimation under each construction provides the corresponding nonlinear response comparison. Neither comparison bounds a distortion shared by all constructions.

Confidence and point-sign breakdown. I distinguish losing confidence evidence for a sign from changing the sign of the estimated response. The confidence-breakdown radius is the smallest M at which the bias-aware 68% pointwise interval includes zero. It is zero if the baseline interval already includes zero. The local point-sign radius divides $|\hat{\beta}_h|$ by the largest absolute linearized response change per unit radius. Neither calculation is an exact nonlinear confidence result.

For a finite perturbation, the response is a ratio. I therefore also calculate when the numerator can reach zero while the denominator stays positive. Using the sample counterpart of the existing support-moment Jacobian, with $R_t = (1, p_t)'$,

$$\hat{G}_h^S = -\text{diag} \left(\frac{1}{n_h} \sum_{t \in \mathcal{T}_h} R_t R'_t, \frac{1}{n_h} \sum_{t \in \mathcal{T}_h} R_t R'_t \right).$$

The fixed affine-moment experiment shifts the four fitted coefficients to

$$\hat{\vartheta}_h(\delta_h) = \hat{\vartheta}_h - (\hat{G}_h^S)^{-1} B_h \delta_h / \sqrt{n_h}, \quad \|\delta_h\|_2 \leq M.$$

Both fitted endpoint moments are linear in this shift. The exact point-sign radius is the smallest $\|\delta_h\|_2$ for which

$$\hat{a}_h(\delta_h) + \hat{b}_h(\delta_h) = 0, \quad \hat{a}_{D,h}(\delta_h) + \hat{b}_{D,h}(\delta_h) > 0.$$

I also report the smallest radius permitting a zero denominator. Because $\hat{G}_h^S$ is block diagonal and B_h is diagonal, numerator and denominator shifts occupy orthogonal coordinate blocks. The minimum-norm numerator-zero shift leaves the denominator unchanged. I nevertheless check its minimum over the entire ball before interpreting the radius. A sign reversal obtained only by making the denominator negative is not an admissible reversal of the response.

Exactness refers to this fixed, four-moment perturbation experiment. It is not an exact coverage statement or a claim that every permitted moment shift can be generated by changing action-level support probabilities. The reported confidence thresholds continue to use the local bias-aware intervals.

C.9 Pooled Applications

The pooled UK–France application allows the numerator support slope to differ across countries while imposing a common high-support numerator endpoint. With a fixed number C of countries, define

$$R_{ct}^P = (1, \mathbf{1}\{c = 1\}(p_{ct} - 1), \dots, \mathbf{1}\{c = C\}(p_{ct} - 1))'. \quad (44)$$

The first coefficient, denoted A_h , is the common numerator value at $p_{ct} = 1$; the remaining coefficients are country-specific support slopes. The denominator is not constrained to have a common endpoint. For each country, I separately project the zero-filled squared residualized VAT change on $(1, p_{ct})$ over all usable quarters and evaluate the fitted relation at $p_{ct} = 1$, obtaining $Q_{c,h}(1)$. The main pooled normalization is

$$\bar{Q}_h(1) = \frac{1}{C} \sum_{c=1}^C Q_{c,h}(1), \quad \hat{\beta}_h^P = \frac{\hat{A}_h}{\hat{\bar{Q}}_h(1)}.$$

An event-count-weighted average of the country denominators is retained as a diagnostic.

For inference, the country-specific outcome and VAT residualization equations, the pooled numerator projection, and the country-specific denominator projections are estimated as one stacked system. Scores at a common calendar quarter are summed before applying the Bartlett-HAC estimator, allowing contemporaneous dependence across countries and serial dependence across quarters. The delta method therefore incorporates residualization uncertainty, numerator–denominator covariance, and cross-country denominator covariance. The bandwidth is $\max\{4, h\}$, as in the U.S. application.

D External Benchmarks and U.S. Empirical Diagnostics

This appendix reports the external-label checks, the harmonized RR comparison, and U.S. measurement diagnostics and an instrument-specific extension under stronger joint restrictions. Online Appendix OA.3 contains supplementary RR documentary and estimator-specific checks. The U.S. fitted $p = 0$ endpoint is reported separately in Online Appendix OA.4.1.

D.1 Romer and Romer replication Robustness checks

D.1.1 Probabilistic validation against the RR classification

As a complementary check, I estimate a simple probabilistic model for whether the original Romer–Romer label is exogenous. Let

$$D_i^{\text{RR}} = \mathbf{1}\{\text{episode } i \text{ is classified as exogenous by Romer–Romer}\}.$$

I fit the one-variable calibration model

$$\Pr(D_i^{\text{RR}} = 1 \mid p_i) = \Lambda\left(a + b \log \frac{p_i}{1 - p_i}\right),$$

where $\Lambda(\cdot)$ is the logistic cdf. The null $b = 0$ corresponds to no monotone predictive content in the policy-level probability. This exercise is deliberately exploratory: the RR classification is used as an external benchmark, not as a definitive latent truth, and the sample contains only 50 policy episodes.

Table 14: Exploratory Logit Validation Against the Original RR Classification

Statistic	Estimate
Number of RR episodes	50
Coefficient on $\log(p/(1 - p))$	0.712
Standard error	0.291
Wald p -value	0.014
Likelihood-ratio p -value	0.006
McFadden pseudo- R^2	0.109
In-sample AUC	0.699
Leave-one-out AUC	0.660

Note: The dependent variable equals one for episodes classified as exogenous by Romer and Romer. The leave-one-out AUC refits the one-variable logit after dropping each episode in turn.

The positive and statistically significant slope indicates that higher policy-level probabilities are systematically associated with the original RR exogenous classification. At the same time, the AUC remains far below one. This is consistent with the interpretation in the main text: the probability contains meaningful narrative information, but ambiguous policy episodes are not perfectly separable by a single continuous score.

D.1.2 Harmonized Binary Selection and Probabilistic Identification

This appendix documents the comparison in Figure 3. The binary benchmark uses a quarterly tax series formed only from tax amounts attached to actions that Romer and Romer classify as exogenous. The probabilistic specification uses the broader quarterly tax series and the continuous policy-level support measure. The two tax series consequently differ by construction. The outcome transformation, lag

controls, timing, horizons, and HAC treatment are common. Binary OLS fits both control projections on every usable quarter, including quarters with zero selected taxes. The probability estimator fits both control projections on recorded-action quarters, sets residual products to zero elsewhere, and projects the affine moments over all usable quarters with no-action support zero. At horizon 8 there are 228 usable quarters and 67 recorded-action quarters. The responses are estimated jointly so that inference for their difference accounts for shared observations, control estimation, and numerator–denominator covariance. HAC standard errors allow for serial dependence in the quarterly moment conditions, with Bartlett bandwidth $\max(4, h)$.

At horizon 8, the binary selected-shock response is -2.25 , compared with -0.78 for the continuous-support response. The probability-minus-binary difference is 1.48 , with a paired 95% interval of $[0.11, 2.84]$, which excludes zero for this specification. The corresponding inflation interval includes zero. These statements condition on constructed support and do not validate the identifying restrictions. The archived comparison that holds the total RR tax series fixed and changes only its binary or continuous support representation is a decomposition diagnostic; it is not the primary economic comparison because it retains tax amounts that the traditional RR design excludes.

Table 15: Romer–Romer GDP Moment Accounting at Eight Quarters

	RR binary (all usable quarters)	Probabilistic
Outcome–tax moment	-0.103	-0.214
Squared-tax moment	0.046	0.276
GDP response (%)	-2.25	-0.78

Table 15 reports the numerator and denominator of the displayed GDP responses. The binary moments average over all usable quarters; the probability moments evaluate their fitted affine relationships at $p = 1$. Under the maintained restrictions, the latter squared-tax moment pertains to the unobserved exogenous type. Its larger fitted value outweighs the larger negative outcome–tax moment. The comparison does not by itself establish that latent exogenous actions are intrinsically larger: its averaging operators and control-fitting samples differ.

D.1.3 Action-Level Probability Construction

This exercise varies the construction of narrative support for recorded RR policy actions while holding the quarterly tax series and macroeconomic specification fixed. The alternatives include the baseline ensemble, evaluator-specific representations, calibration-family averages, individual calibration samples, and PCA reductions. They are alternative constructions of the same generated support measure, not independent replications. At horizon 8, all 22 constructions yield a negative GDP response that is smaller in absolute value than the RR binary selected-shock response. The ordering is less uniform away from the central medium-run horizon: at horizons 4, 12, and 20, the number of negative constructions is respectively 16, 21, and 21 out of 22. Figure 11 summarizes the resulting range of GDP responses. The separate no-action completion exercise in Appendix F changes the treatment of quarters without recorded policy actions after quarterly aggregation; it does not alter the construction of support for recorded RR actions.

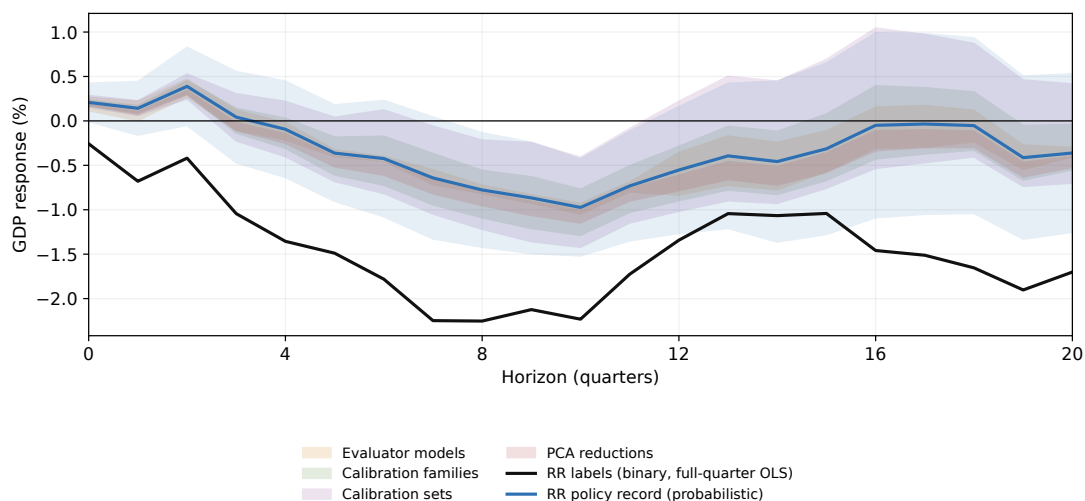


Figure 11: Romer–Romer GDP Response Across Action-Level Support Constructions

Note: The black line reports the RR binary response and the thick blue line the probabilistic response. The blue band is the baseline 68% HAC interval. Colored envelopes summarize the responses across 22 probability constructions, covering evaluator models, individual calibration sets and PCA reductions.

D.1.4 Binary Inclusion Threshold Diagnostic

This exercise is diagnostic. The baseline probabilistic framework does not require a support threshold; the results below instead show how the estimated response changes when continuous narrative support is deliberately collapsed into a binary inclusion rule. Figure 12 reports controlled OLS on all usable quarters for each cutoff. A cutoff retains the whole quarterly net tax change when absolute-tax-weighted quarterly support is at least $\bar{p}$, setting the raw selected series to zero otherwise. Original RR labels instead select individual action amounts before quarterly aggregation.

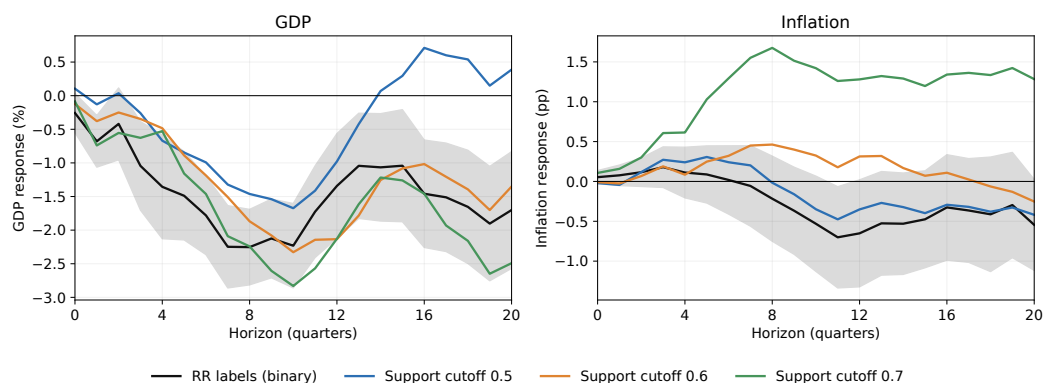


Figure 12: Romer–Romer Responses under Binary Support Cutoffs

Note: Every curve uses controlled OLS on all usable quarters and the same lag controls. Cutoffs 0.5, 0.6, and 0.7 retain the whole quarterly tax change using the same measured support. Original RR labels select action amounts before aggregation. The shaded 68% HAC band is shown only for the original-label benchmark. No continuous-support curve is displayed in this diagnostic.

At horizon 8, the GDP estimates are -1.46 , -1.87 , and -2.24 for cutoffs 0.5, 0.6, and 0.7, respectively, compared with -2.25 under original RR selection. The cutoff rules retain 44, 35, and 20 quarters with nonzero selected tax changes, while all four OLS fits use 228 usable quarters. Inflation responses are -0.02 , 0.46 , and 1.68 percentage points, respectively, compared with a decline of 0.22 under original labels. Thus, the 0.7 GDP estimate is close to the original-label benchmark, but the magnitude of the contraction and the inflation response depend on the inclusion rule. These changes reflect different selected policy compositions, not progressively cleaner versions of a fixed response. Online Appendix OA.3.4 discusses the associated narrative ambiguities.

D.1.5 Probabilistic validation against the Cloyne classification

I apply the same one-variable probabilistic check to the Cloyne validation sample. Let

$$C_i = \mathbf{1}\{\text{policy row } i \text{ is classified as exogenous by Cloyne}\}.$$

I estimate

$$\Pr(C_i = 1 \mid p_i) = \Lambda\left(a + b \log \frac{p_i}{1 - p_i}\right),$$

where the sample consists of Cloyne policy rows with usable canonical labels and complete probability inputs. Because many policy rows belong to the same narrative event, I report both row-level and event-clustered uncertainty, and I also evaluate prediction after holding out entire narrative events.

Table 16: Exploratory Logit Validation Against the Cloyne Classification

Statistic	Estimate
Number of Cloyne policy rows	2,101
Coefficient on $\log(p/(1 - p))$	0.579
Row-level standard error	0.048
Row-level Wald p -value	< 0.001
Event-clustered standard error	0.087
Event-clustered Wald p -value	< 0.001
Likelihood-ratio p -value	< 0.001
McFadden pseudo- R^2	0.059
In-sample AUC	0.655
Leave-one-event-out AUC	0.641

Note: The dependent variable equals one for policy rows classified as exogenous by Cloyne. Event-clustered inference groups policy rows by narrative event. The held-out AUC refits the one-variable logit after dropping each event in turn.

The Cloyne probability is therefore statistically predictive of the external expert label: the slope is positive and remains precisely estimated after clustering by narrative event. The predictive strength is still modest, however. The AUC is 0.65 in sample and 0.64 when entire narrative events are held out. This is consistent with the main-text table: the probability ranks Cloyne-exogenous rows above Cloyne-endogenous rows on average, but the two distributions overlap substantially. I interpret this as evidence of monotone external-label validity, not as evidence that the procedure mechanically reproduces Cloyne’s binary taxonomy.

D.2 Measurement Diagnostics for the U.S. Budget Endpoint

This appendix reports construction-sensitivity comparisons, classical corrected-moment benchmarks under additional measurement assumptions, and breakdown thresholds for the approximate-moment stress test. It relates permitted distortions to the magnitude and direction of observed measurement changes.

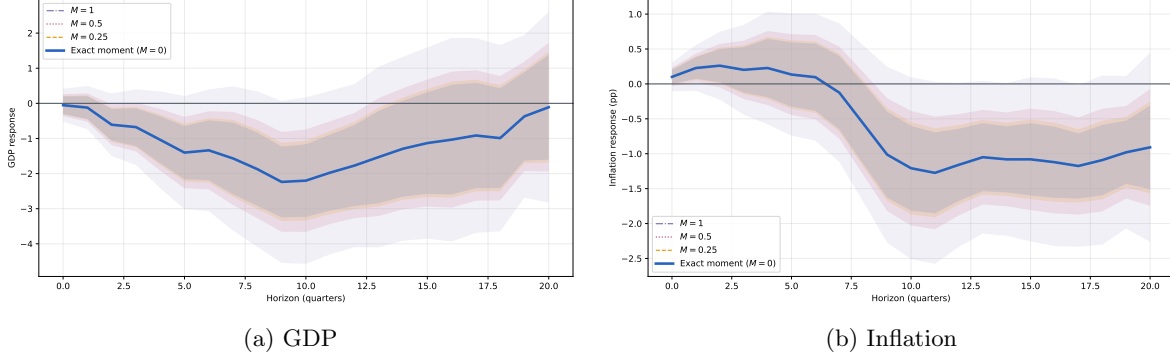


Figure 13: U.S. Budget Endpoints: Approximate-Moment Sensitivity

Note: Each panel reports 68% pointwise bias-aware intervals for the high-support response. $M = 0$ reproduces the baseline full-calendar estimate and its fully stacked HAC interval. Larger values allow deviations of the four affine moments, scaled by their nuisance-profiled HAC variability. The permitted response bias is evaluated locally around the baseline; the sampling variance and moment scales are held fixed. These are not simultaneous bands or exact nonlinear ratio intervals.

Breakdown and denominator checks. Table 17 separates confidence breakdown from a change in the point response’s sign. At horizons 4, 8 and 12, the GDP intervals exclude zero at 68%, but lose this sign conclusion before the permitted distortion can bring the point estimate to zero. Inflation’s baseline 68% intervals already include zero at horizons 4 and 8; only horizon 12 has a positive confidence-breakdown radius.

The fitted denominator remains positive throughout each ball up to the exact point-sign radius. Separately, each of the 2,160 constructions has a positive denominator and a positive pointwise 95% lower confidence limit at all three horizons. These denominators are common to GDP and inflation. No construction is excluded by these warning checks, which do not constitute a general weak-identification test.

Figure 14 shows that the model-specific endpoints are not identical, but they preserve similar medium-run patterns for both GDP and inflation. This diagnostic indicates that the main pattern is not driven by a single representation model.

Figure 15 shows that varying the motivations extracted from the same retrieved passages changes the magnitude of some estimates but preserves the contractionary GDP response and the medium-run decline in inflation. The draw-averaged extraction-model probabilities have a mean Pearson correlation of 0.80 with one another; stochastic draws within an extraction model have a mean correlation of 0.93.

Figure 16 provides reference magnitudes for interpreting the sensitivity radius under one-at-a-time perturbations. Across the main horizons, calibration-sample perturbations have median equivalent radii around $M = 0.5$, representation-model perturbations around $M = 0.7$, and motivation-extraction

Table 17: Confidence and point-sign breakdown in the four-moment experiment

Outcome	h	Baseline response	68% CI radius	Local sign radius	Exact sign radius	Denominator zero radius
GDP	4	-1.039	0.673	0.976	1.037	2.888
GDP	8	-1.874	0.801	1.087	1.187	2.708
GDP	12	-1.772	0.680	0.993	1.073	2.618
Inflation	4	0.228	0.000	0.372	0.376	2.888
Inflation	8	-0.568	0.000	0.671	0.692	2.708
Inflation	12	-1.159	0.982	1.243	1.412	2.618

Note: Radii use the fixed nuisance-profiled four-moment scale. Zero confidence radius means the baseline pointwise 68% interval already includes zero. Confidence and local sign calculations use the derivative at the baseline. The exact sign radius attains a zero numerator with positive denominator in the fixed affine-coefficient experiment; it is not an exact coverage result. The entire ball remains denominator-positive through each reported exact sign radius. Larger radii can permit a sign reversal. Baseline response units are percent for GDP and percentage points for inflation.

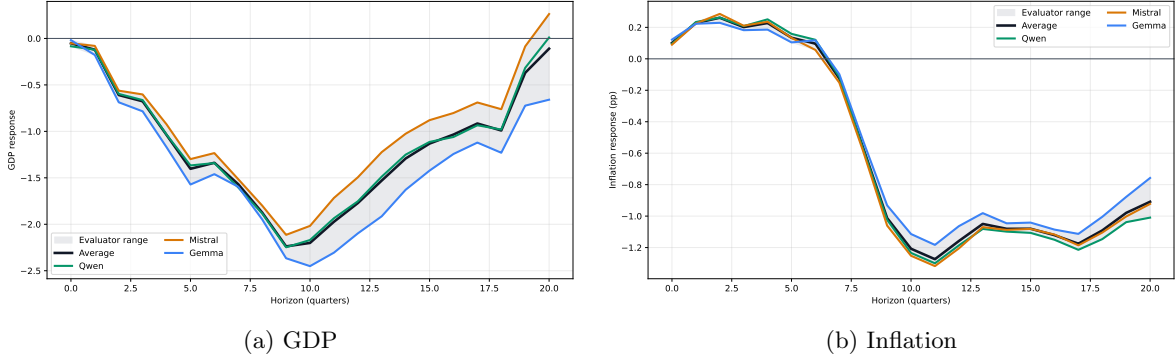


Figure 14: U.S. Budget Endpoints: Representation-Model Stability

Note: Each line recomputes the same endpoint using support scores from one representation-model family. The exercise isolates sensitivity to the language-model representation used to construct the support score.

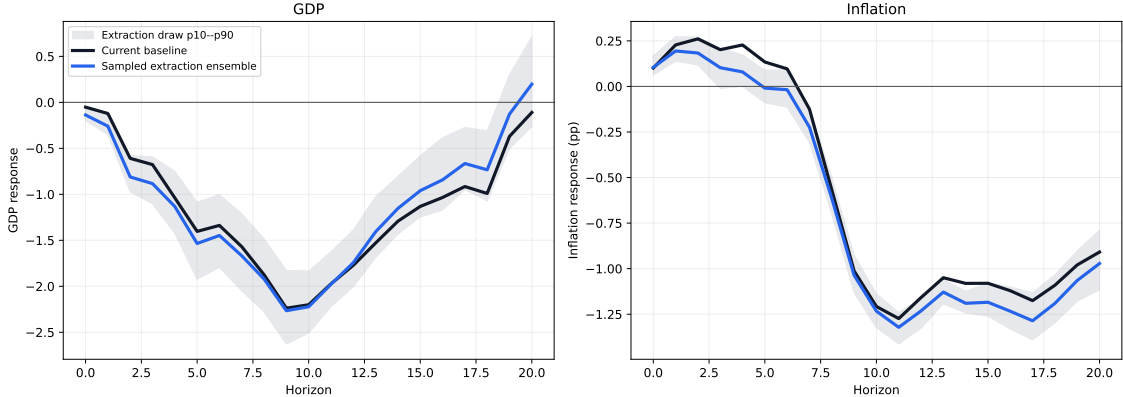


Figure 15: U.S. Budget Endpoints: Motivation-Extraction Sensitivity

Note: The black line uses the baseline support score. The blue line uses the support score averaged across alternative motivation extractions. The gray region is the 10th–90th percentile range of endpoint estimates across 60 constructions: 20 stochastic motivation draws from each of Qwen-9B, Gemma-4B, and Mistral-7B. Retrieval is held fixed, and each extraction is passed through the same evaluator-model and calibration-set construction.

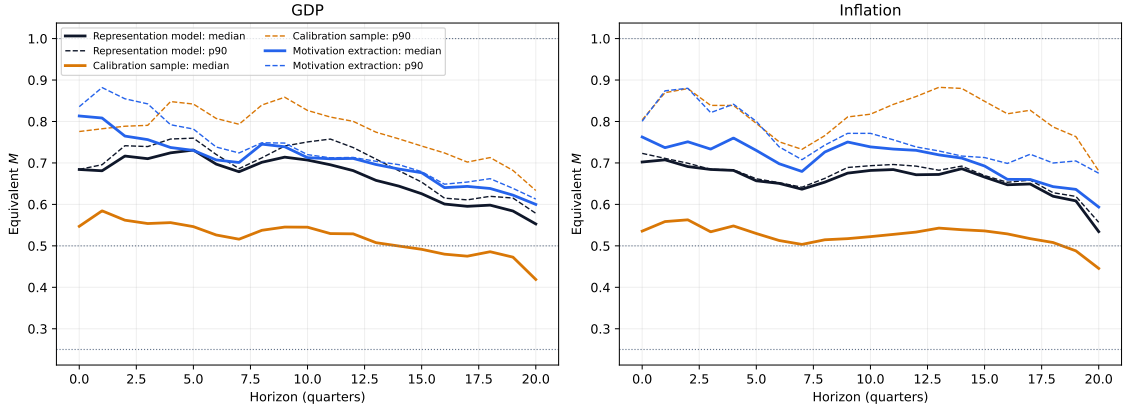


Figure 16: Observed Probability Perturbations in Equivalent Sensitivity Units

Note: For each alternative evaluator model, calibration set, or motivation-extraction construction, the figure evaluates the four affine moments at the baseline parameter and replaces only the support sequence. The resulting moment displacement is scaled by the nuisance-profiled HAC variability used in Figure 13. Solid lines report the median equivalent M within each family; dashed lines report the 90th percentile.

perturbations around $M = 0.7$. These values should not be interpreted as estimates of structural misspecification: they report how far the implemented moments move when observed measurement choices are changed.

Observed directions and the stress test. At horizons 4, 8 and 12, median GDP directional alignment among the joint constructions is 0.269, 0.187 and 0.177. Their median, 0.19, is the summary reported in the Table 18; the inflation summary is also 0.19. This limited alignment helps explain why observed constructions preserve negative GDP responses even when their equivalent radii exceed worst-direction breakdown thresholds. Alignment is a local comparison; the re-estimated responses provide the nonlinear check. Neither comparison bounds shared error.

Classical corrected-moment benchmarks. Table 19 separates the effect of changing the measurement average from the additional classical correction. With the baseline 36 constructions, the GDP response at horizon 8 changes from -1.87424 to -1.87405 after correction. At this horizon, changing the measurement average has a larger effect than the additional correction. Over both outcomes and horizons 0–20, the largest absolute corrections are 0.004675 for the baseline average and 0.002450 for the expanded average.

The small ratio corrections do not imply that every moment is unchanged: numerator and denominator changes can offset. Nor do they establish that shared measurement error is small. These comparisons describe the consequences of the stipulated classical model, not a correction validated against observed latent support.

Table 18: Magnitude and Endpoint Relevance of Observed Support Perturbations

Outcome	Source	Median M^{eq}	P90 M^{eq}	Median relevance	Median $ \Delta\hat{\beta} $
GDP	Representation model	0.70	0.74	0.15	0.114
GDP	Calibration sample	0.54	0.84	0.11	0.098
GDP	Extraction model	0.74	0.75	0.27	0.256
GDP	Stochastic extraction draw	0.09	0.16	0.56	0.067
GDP	All choices jointly	0.85	2.20	0.19	0.267
Inflation	Representation model	0.67	0.68	0.07	0.024
Inflation	Calibration sample	0.53	0.84	0.19	0.097
Inflation	Extraction model	0.73	0.74	0.15	0.060
Inflation	Stochastic extraction draw	0.09	0.16	0.58	0.032
Inflation	All choices jointly	0.82	2.24	0.19	0.144

Note: Each entry is the median of the corresponding horizon-specific statistic at horizons 4, 8 and 12. M^{eq} is the standardized size of the observed displacement of the four support moments, using their nuisance-profiled HAC scales. Relevance is the absolute linearized response change divided by the largest absolute linearized change at the same radius. The final column reports the absolute response change after re-estimation. Stochastic draws are compared with their extraction-model draw average; other constructions are compared with the baseline. The separate summary statistics need not describe the same construction. They provide magnitude and direction references, not bounds on shared measurement error.

Table 19: Classical benchmarks: changing the measurement average versus correcting it

Outcome	h	Baseline average (36)		Expanded average (2,160)	
		Uncorrected	Corrected	Uncorrected	Corrected
<i>Panel A: point estimates</i>					
GDP	4	-1.03857	-1.03838	-1.13245	-1.13301
GDP	8	-1.87424	-1.87405	-1.92083	-1.92102
GDP	12	-1.77197	-1.77285	-1.74337	-1.74397
Inflation	4	0.22755	0.22831	0.07961	0.07914
Inflation	8	-0.56763	-0.56980	-0.61385	-0.61625
Inflation	12	-1.15850	-1.16000	-1.23171	-1.23369
<i>Panel B: pointwise 68% intervals</i>					
GDP	4	[-1.706, -0.371]	[-1.706, -0.371]	[-1.754, -0.511]	[-1.755, -0.511]
GDP	8	[-2.918, -0.831]	[-2.919, -0.829]	[-2.823, -1.019]	[-2.824, -1.018]
GDP	12	[-2.925, -0.619]	[-2.929, -0.617]	[-2.745, -0.742]	[-2.747, -0.741]
Inflation	4	[-0.190, 0.645]	[-0.190, 0.647]	[-0.312, 0.471]	[-0.313, 0.471]
Inflation	8	[-1.148, 0.012]	[-1.150, 0.010]	[-1.169, -0.059]	[-1.171, -0.062]
Inflation	12	[-1.675, -0.642]	[-1.677, -0.643]	[-1.749, -0.714]	[-1.751, -0.716]

Note: GDP responses are percentages and inflation responses are percentage points. The corrected columns impose the stated classical repeated-measurement restrictions. The expanded average includes extraction variation and is not the paper baseline. Intervals use the full stacked covariance, including control estimation, the two endpoint moments and the local correction statistic where applicable. The 2,160 constructions are not independent replications. These intervals do not test differences between columns.

D.3 Instrument-Specific Responses under Stronger Joint Restrictions

The decomposition in Table 6 identifies additive contributions to the aggregate budget response, not responses holding other tax instruments fixed. Separating the latter matters because tax groups frequently move together: 88 of the 134 quarters with recorded actions contain more than one group. Presence is measured from gross action amounts, so offsetting actions remain visible. This is evidence of quarterly co-occurrence, not necessarily of a common legislative package.

The instrument-specific object. Let $\widetilde{\Delta\tau}_{t,h}$ collect the residualized quarterly changes in the five tax groups used in the main decomposition, with the same zero-filling convention as Subsection 4.2. Consider a common linear, additive response equation

$$\widetilde{y}_{t,h} = \sum_{k=1}^5 \beta_{k,h}(1) \widetilde{\Delta\tau}_{k,t,h} + \varepsilon_{t,h}.$$

Here $\varepsilon_{t,h}$ is the disturbance in this joint equation, rather than the disturbance in the scalar aggregate-tax equation. Each coefficient describes the response to an increase in one tax group holding the other contemporaneous groups fixed. Collect the conditional outcome-tax moments and tax second moments as

$$\begin{aligned} N_h(p) &= E\left[\widetilde{\Delta\tau}_{t,h} \widetilde{y}_{t,h} \mid p_t = p\right], \\ Q_h(p) &= E\left[\widetilde{\Delta\tau}_{t,h} \widetilde{\Delta\tau}_{t,h}' \mid p_t = p\right]. \end{aligned}$$

The diagonal of $Q_h(p)$ contains own-tax squared moments; its off-diagonal entries contain cross-tax moments. Unlike the common scalar normalization in the additive decomposition, this matrix allows the point estimates to account for taxes moving together.

Additional identifying restrictions. Four restrictions are maintained for an instrument-specific causal interpretation. First, the linear, additive equation must provide a common partial-response vector; omitted tax interactions or package-specific responses need not have this interpretation. Second, the support restrictions must recover the relevant high-support endpoint for every entry of $N_h(p)$ and $Q_h(p)$. The empirical implementation fits each entry as affine in the same aggregate support index. The stability requirement therefore extends to all outcome-tax and own- and cross-tax moments: support must not sort within-type tax size, signs or composition in ways that change these moments. Action-level calibration and quarterly support aggregation alone do not establish this extension. Nor is aggregate support interpreted as the probability that every tax in a quarter is exogenous. An endpoint outside observed support relies on the maintained affine extension rather than an observed group with $p_t = 1$.

Third, exclusion must hold separately for every tax group, so that

$$E\left[\widetilde{\Delta\tau}_{t,h} \varepsilon_{t,h} \mid p_t = 1\right] = 0.$$

Aggregate orthogonality would only restrict the sum of these instrument-disturbance moments and could conceal offsetting biases. Fourth, $Q_h(1)$ must have full rank: fixed-proportion tax packages

cannot identify separate responses. Under these restrictions, $N_h(1) = Q_h(1)\beta_h(1)$, giving the estimator

$$\hat{\beta}_h(1) = \hat{Q}_h(1)^{-1} \hat{N}_h(1).$$

Inverting the matrix does not remove correlated spending changes, anticipation or other omitted policy disturbances. The controls retain aggregate tax lags, not a separate history for each instrument.

Estimation and results. The exercise preserves the baseline U.S. sample, probabilities and controls. Control projections are fitted on quarters with nonzero aggregate net tax variation, and residuals are zero-filled elsewhere. All usable quarters enter the affine projections; tax changes and support remain zero in no-action quarters. Five outcome–tax products and fifteen distinct own- and cross-tax products are projected on $(1, p_t)$ and evaluated at $p = 1$. Fully stacked HAC standard errors allow for serial dependence in the quarterly moment conditions, using Bartlett bandwidth $\max\{4, h\}$. The matrix-inverse delta method includes uncertainty and covariance from the control projections, numerator vector and entire second-moment matrix, following the inference construction in Appendix C.

All 63 outcome–horizon fits yield positive-definite endpoint matrices, with diagonal-standardized condition numbers between 3.58 and 3.81. These are numerical admissibility checks, not tests of exclusion, stability or statistical identification strength. The pointwise intervals are not weak-identification-robust and condition on the constructed support sequence and fixed expenditure shares.

Table 20: Instrument-Specific Responses at Eight Quarters

Tax group	GDP	Consumption	Investment
Individual income	−1.49 (1.03)	−1.21 (0.64)	−1.03 (0.58)
Corporate income	−2.13 (2.96)	−0.80 (1.80)	−1.32 (1.56)
Selective excises and bundles	−4.48 (4.33)	−1.02 (2.18)	−0.99 (1.88)
Energy, transport and other indirect taxes	4.54 (1.90)	3.01 (1.35)	1.96 (1.00)
Other low-support taxes	2.88 (2.95)	1.30 (1.86)	0.53 (1.20)

Note: Responses to a proposed revenue increase equivalent to one percent of GDP in the indicated tax group, holding the other contemporaneous groups fixed under the restrictions in this appendix. Parentheses contain fully stacked HAC standard errors, including control projections and all numerator and cross-tax moments. Consumption and investment are converted to approximate percentage points of GDP using fixed sample-mean nominal expenditure shares of 0.6250 and 0.1746; GDP is in percent. These are instrument-specific responses, not additive contributions to the aggregate budget response. Inference conditions on the constructed support sequence and expenditure shares.

At eight quarters, individual-income estimates are negative for GDP, consumption and investment. The corporate-income GDP estimate is also negative, but imprecise. Indirect-tax results remain heterogeneous, including positive estimates for the broader energy, transport and other indirect group. These are partial responses under stronger maintained restrictions, not additive contributions or validation of causal assumptions. Online Appendix OA.4.3 reports the complete response paths.

D.4 Binary Selection and Instrument-Level Responses

Tables 21 and 22 compare eight-quarter GDP responses using the same measured support continuously or converted to binary selection at cutoffs 0.6, 0.7 and 0.8. Selection retains individual tax actions before quarterly aggregation. Controls follow Section 4.2 but are fitted on quarters with nonzero selected net tax changes in the binary columns. Selection changes the policy mix, fitted controls and normalization.

Table 21: Additive GDP Contributions under Binary Selection

Tax group	Probabilistic	Binary cutoff		
		0.6	0.7	0.8
All-budget total	−1.87** (1.05)	−2.44** (1.39)	−3.29** (1.53)	−2.36* (1.58)
Individual income	−2.01** (0.96)	−2.17** (1.26)	−2.74** (1.18)	−1.83* (1.50)
Corporate income	−0.86* (0.53)	−1.71* (1.14)	−0.53* (0.52)	−0.26 (0.81)
Selective excises and bundles	−0.24* (0.19)	−0.33* (0.27)	−0.04 (0.13)	−0.23* (0.14)
Energy, transport and other indirect taxes	0.91* (0.56)	1.87* (1.35)	0.06 (0.13)	−0.05* (0.04)
Other low-support taxes	0.32* (0.25)	−0.12* (0.10)	−0.05 (0.09)	0.01 (0.03)
Control-fit quarters	129	96	76	60

Note: Eight-quarter GDP contributions for a proposed tax increase of one percent of GDP. Entries sum to their column total, with column-specific aggregate normalization. Grey parentheses contain fully stacked HAC standard errors. * and ** indicate that pointwise 68% and 90% intervals exclude zero, respectively.

At cutoff 0.7, the aggregate GDP response is −3.29 percent, compared with −1.87 under probabilistic identification. The selected-record contraction is concentrated in individual income taxes, with indirect-tax contributions close to zero. Additive magnitudes and precision reflect their place in the selected mix.

Table 22: Instrument-Specific GDP Responses under Binary Selection

Tax group	Probabilistic	Binary cutoff		
		0.6	0.7	0.8
Individual income	−1.49* (1.03)	−1.31 (1.66)	−3.69** (1.93)	−0.91 (1.36)
Corporate income	−2.13 (2.96)	−3.05 (4.57)	−4.37 (6.10)	−2.03 (6.97)
Selective excises and bundles	−4.48* (4.33)	−9.36 (10.68)	1.62 (7.46)	−72.74** (24.25)
Energy, transport and other indirect taxes	4.54** (1.90)	6.70** (3.22)	3.55 (3.91)	−190.17** (69.85)
Other low-support taxes	2.88 (2.95)	−12.68 (17.13)	1.55 (20.77)	20.67 (22.06)
Control-fit quarters	129	96	76	60

Note: Eight-quarter percent GDP responses to a proposed revenue increase of one percent of GDP in the indicated tax group, holding the other included tax groups fixed under the joint restrictions in Appendix D.3. Grey parentheses contain fully stacked HAC standard errors. * and ** indicate that pointwise 68% and 90% intervals exclude zero, respectively.

All five tax groups remain jointly estimable under the restrictions in Appendix D.3. At cutoff 0.7, corporate-income and selective-excise standard errors rise from 2.96 and 4.33 to 6.10 and 7.46. At cutoff 0.8, the large indirect-tax coefficients use very little surviving variation.

Causal interpretation requires each specification’s exclusion restrictions, including restrictions on correlated effects of excluded contemporaneous policies.

E VAT Shocks and Macro Transmission in the United Kingdom and France

E.1 Sensitivity to Action-Level Support Construction

The baseline support measure averages across representation models and calibration samples. To isolate sensitivity to these action-level measurement choices, I reestimate the pooled VAT application using 22 constructions: the baseline ensemble, three representation-model averages, four calibration-family averages, twelve individual calibration samples, and two PCA reductions. Tax amounts, the 1955Q1–2007Q4 sample, quarterly aggregation, controls, residualization, denominator specification, and inference procedure remain fixed; both numerator and denominator moments are reestimated under each support construction. Quarters without recorded VAT actions continue to enter with support equal to zero, so this exercise is distinct from the quarterly-completion analysis in Appendix F. Because the constructions share documents and calibration components, their dispersion is interpreted as measurement sensitivity rather than sampling uncertainty.

Table 23: VAT Sensitivity to Action-Level Support Construction

Outcome	h	Baseline	Minimum	Median	Maximum	Sign retained
GDP	4	-2.46	-2.68	-2.44	-2.31	22/22
	8	-4.55	-4.83	-4.55	-4.26	22/22
	12	-4.04	-4.26	-4.04	-3.64	22/22
Inflation	4	4.26	3.48	4.25	5.25	22/22
	8	4.54	3.65	4.54	5.08	22/22
	12	2.11	1.77	2.11	2.36	22/22
Private consumption contribution	4	-1.61	-1.72	-1.61	-1.50	22/22
	8	-2.47	-2.57	-2.46	-2.31	22/22
	12	-3.01	-3.18	-3.01	-2.80	22/22
Investment contribution	4	-0.56	-0.61	-0.56	-0.51	22/22
	8	-1.30	-1.42	-1.30	-1.21	22/22
	12	-1.23	-1.31	-1.24	-1.11	22/22

Note: The table compares 22 action-level narrative-support constructions: the baseline ensemble, three representation-model averages, four calibration-family averages, twelve individual calibration samples, and two PCA reductions. Minimum, median, and maximum are calculated across the 22 estimates. “Sign retained” reports the number retaining the baseline sign. Tax amounts, the sample, quarterly aggregation, controls, residualization, denominator specification, and inference procedure are held fixed; both numerator and denominator moments are reestimated under each support construction. GDP, private consumption, and investment are changes from the pre-event quarter in percentage points of trend GDP; inflation is in percentage points. The constructions share documents and calibration components and are not independent replications.

Online Appendix Figure OA.14 reports the complete paths. At horizons 4, 8, and 12, all 22 constructions preserve the negative GDP response, positive inflation response, and negative private-consumption and investment contributions. At horizon 8, GDP ranges from -4.83 to -4.26 , inflation from 3.65 to 5.08 , the private-consumption contribution from -2.57 to -2.31 , and the investment contribution from -1.42 to -1.21 . Differences in sign arise only near the end of the GDP and investment paths, where the corresponding baseline responses are close to zero.

E.2 VAT Packages and the Nguyen–Onnis–Rossi Concern

Nguyen et al. (2021) emphasize that UK income-tax and consumption-tax changes frequently occur together. In their proxy-SVAR, the broad consumption-tax object is an average consumption tax rate built from VAT and consumption duties. They find modest effects for this broad object once income-tax and consumption-tax shocks are jointly identified, while single-proxy specifications can make consumption taxes look much more powerful. This appendix asks whether the same concern jeopardizes the VAT endpoint reported in the main text.

The checks keep the UK–France sample fixed at 1955Q1–2007Q4. They vary only the tax actions entering the VAT object. The first diagnostic is a co-movement audit. Same-quarter overlap is high, but the VAT-income correlation is small: the pooled correlation between the VAT shock and the income-like shock is 0.02 over all quarters and 0.04 over policy quarters. An explicit two-instrument denominator diagnostic gives nearly identical GDP endpoints. At horizons 4, 8, and 12, the naive VAT endpoints are -2.39 , -4.43 , and -3.92 , while the VAT/income adjusted endpoints are -2.40 , -4.44 , and -3.94 .

Table 24 reports the main sensitivity estimates. Dropping UK 1979Q3, the VAT unification episode highlighted by Nguyen et al. (2021), weakens precision but preserves the negative GDP response. Dropping both UK 1973Q2 and UK 1979Q3 weakens the result more substantially, which shows that large UK VAT transition episodes are empirically important. By contrast, dropping VAT actions in quarters where income-like taxes move in the opposite direction and are at least half as large as the VAT move makes the response stronger, not weaker. Thus the VAT result is not driven by compensating VAT-income packages. The main attenuation occurs when the VAT object is broadened to include duties and other excise-like measures, closer to the Nguyen–Onnis–Rossi average-consumption-tax-rate object.

Table 24: VAT Sensitivity to Tax Packages and Broad Consumption-Tax Definitions

Specification	$h = 4$	$h = 8$	$h = 12$
Baseline VAT	-2.46 (< 0.001)	-4.55 (< 0.001)	-4.04 (< 0.001)
Drop UK 1979Q3 VAT	-2.32 (< 0.001)	-3.45 (< 0.001)	-2.80 (0.053)
Drop UK 1973Q2 and 1979Q3 VAT	-1.72 (0.073)	-3.46 (0.006)	-4.44 (< 0.001)
Drop compensating VAT–income packages	-3.02 (< 0.001)	-5.08 (< 0.001)	-4.92 (< 0.001)
Drop full compensating-package quarters	-3.06 (< 0.001)	-5.17 (< 0.001)	-4.76 (< 0.001)
VAT plus duties/excise-like measures	-0.69 (0.037)	-1.57 (0.009)	-1.57 (0.026)

Note: Entries are pooled GDP endpoints at $p = 1$, in percentage points of trend GDP; p-values from the fully stacked HAC covariance are in parentheses. The sample is 1955Q1–2007Q4 in every row. The broad row combines VAT with duties and excise-like measures.

Complete package-sensitivity paths and country-specific diagnostics appear in Online Appendix Figure OA.15 and Online Appendix OA.6.4, respectively.

F Treatment of Quarters without Recorded Policy Actions

The baseline specification retains all calendar quarters and assigns $\Delta\tau_t = p_t = 0$ when no policy action is recorded. The tax change is then observed to be zero, while the value assigned to narrative support is a completion used for estimation rather than an economic support value. This appendix assesses the empirical relevance of that completion.

Let

$$I_t = \mathbb{1}\{\text{at least one policy action is recorded in quarter } t\}.$$

For $c \in \{0, 0.1, \dots, 1\}$, define

$$p_t(c) = \begin{cases} p_t, & I_t = 1, \\ c, & I_t = 0. \end{cases}$$

The value c is a sensitivity parameter, not an estimate of narrative support in quarters without a policy action. The indicator is based on recorded policy rows, so quarters with offsetting actions and a zero net tax change satisfy $I_t = 1$ and retain their measured support. For each value of c , I re-estimate the affine numerator and denominator projections while holding the tax changes, support in recorded-action quarters, controls, horizons, and outcome transformations fixed. A second specification retains the same controls, transformations, and residualization procedure but estimates both the residualization equations and the affine numerator and denominator projections using recorded-action quarters only. The original quarterly spacing is retained when constructing HAC standard errors.

F.1 Main U.S. application

Figure 17 reports the U.S. GDP and inflation responses at horizons 4, 8, and 12. GDP remains negative at every value of c and under the recorded-action-quarter specification. The medium-run inflation response also retains its baseline sign. The estimates can change in magnitude as the no-action mass is moved along the support scale, but the contractionary output response and medium-run disinflationary pattern are preserved.

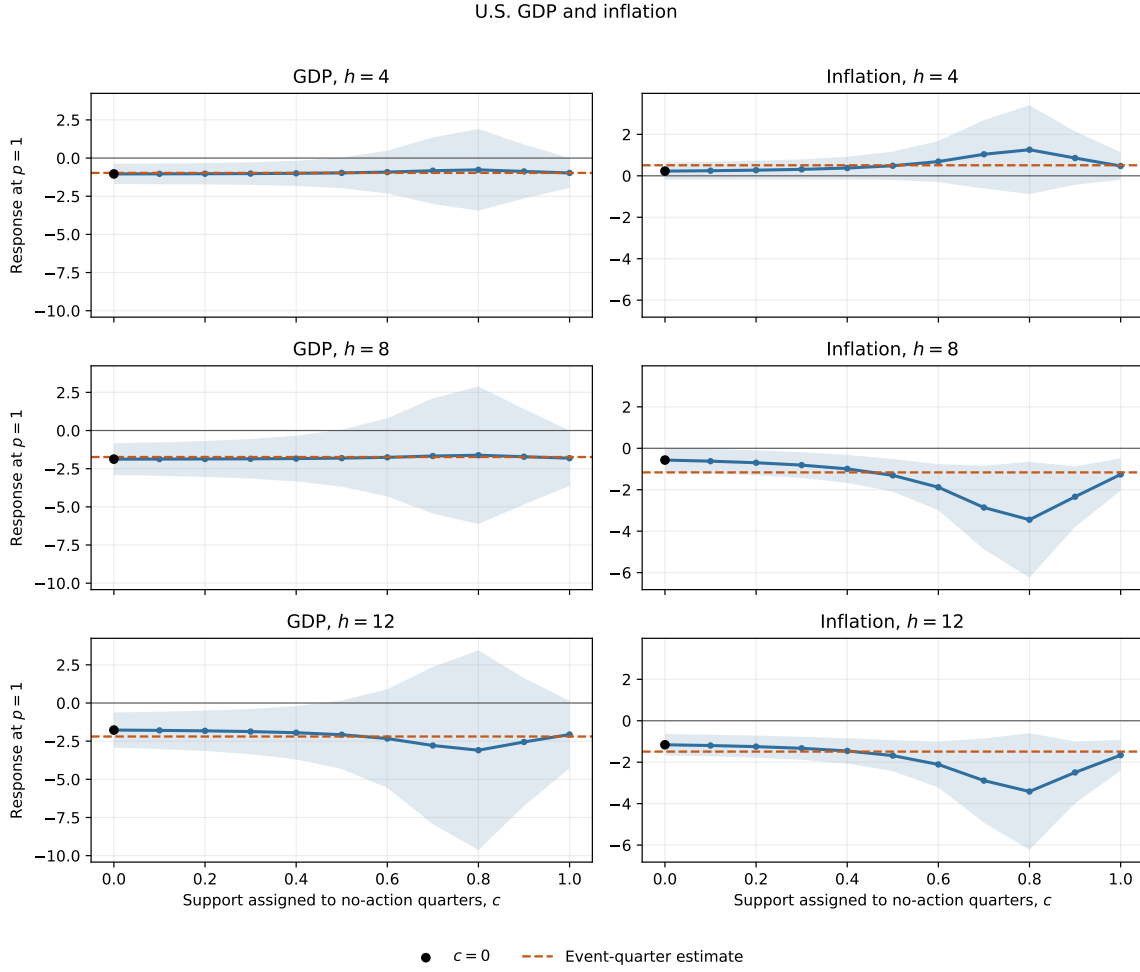


Figure 17: U.S. Responses under Alternative Placements of No-Action Quarters

Note: Each panel reports the response evaluated at $p = 1$ after assigning the common sensitivity value c to quarters without a recorded budget action. The black marker is the baseline $c = 0$ estimate. The orange dashed line is the estimate using recorded-action quarters only. Shaded areas are pointwise 68% fully stacked HAC confidence bands.

Table 25 reports the principal mechanism variables under selected completions and recorded-action-quarter estimation. U.S. consumption is negative in each displayed specification. The investment contraction remains large at horizons 8 and 12 throughout the full c -grid, supporting the large proportional investment response reported in the main text.

Table 25: Mechanism Responses under Alternative Treatments of No-Action Quarters

Application	Outcome and horizon	$c = 0$	$c = 0.5$	$c = 1$	Recorded actions
U.S. budget	Consumption, $h = 4$	-1.42	-1.32	-1.32	-1.45
	Consumption, $h = 8$	-1.87	-1.45	-1.47	-1.53
	Consumption, $h = 12$	-1.22	-0.78	-0.80	-0.92
	Investment, $h = 4$	-3.52	-2.63	-2.67	-2.60
	Investment, $h = 8$	-6.50	-8.02	-7.93	-7.55
	Investment, $h = 12$	-6.05	-9.99	-9.82	-10.41
UK–France VAT	Private consumption, $h = 4$	-1.61	-1.87	-0.68	-1.08
	Private consumption, $h = 8$	-2.47	-2.83	-1.02	-1.61
	Private consumption, $h = 12$	-3.01	-3.41	-1.23	-1.90

Note: The table reports responses evaluated at $p = 1$. “Recorded actions” restricts both first-step residualization and the support-indexed projections to quarters containing at least one recorded policy action while preserving calendar spacing for HAC inference. For the pooled VAT estimates, the fitted country-specific endpoint denominators fail the maintained positivity condition at $c = 0.9$, so no pooled response is reported at that value.

F.2 Pooled VAT application and supporting diagnostics

Figure 18 reports the corresponding pooled UK–France VAT paths. The recorded-action-quarter estimates and the completions $c = 0.5$ and $c = 1$ preserve the output contraction and the initial rise in inflation. Table 25 shows that private consumption also remains negative. Magnitudes are more sensitive than in the U.S. application, and the fitted country-specific endpoint denominators fail the maintained positivity condition at $c = 0.9$. The pooled VAT evidence therefore supports the qualitative mechanism, while also showing that its normalization is less stable under some high-support completions.

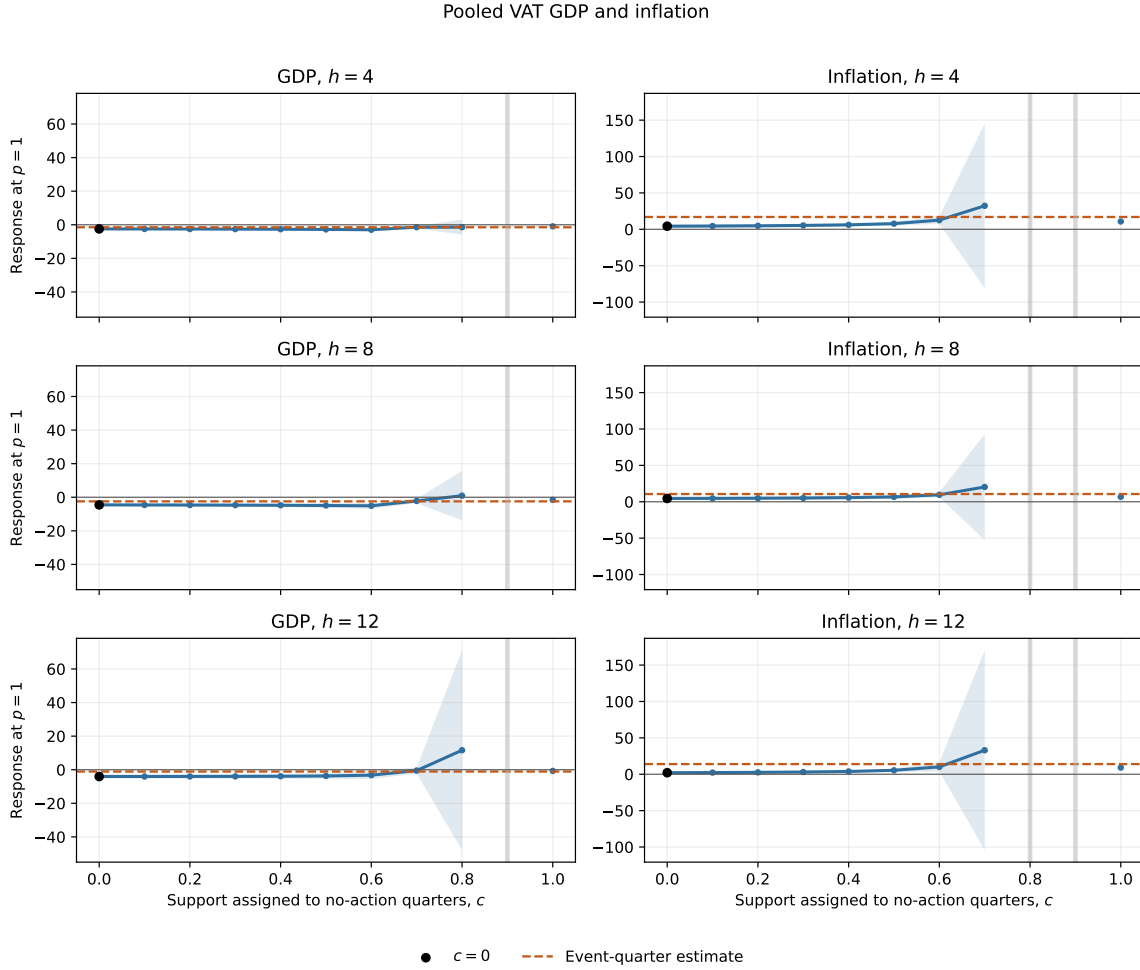


Figure 18: Pooled VAT Responses under Alternative Placements of No-Action Quarters

Note: Each panel reports the pooled UK–France response evaluated at $p = 1$ after assigning the common sensitivity value c to country-quarters without a recorded VAT action. The black marker is the baseline $c = 0$ estimate, and the orange dashed line is the estimate using recorded-action quarters only. Shaded areas are pointwise 68% fully stacked HAC confidence bands. Grey markers identify values of c for which the fitted endpoint denominator fails the maintained positivity condition.

The sparse Romer–Romer and France-only exercises do not provide the same denominator stability. In the Romer–Romer sample, the fitted endpoint denominator is nonpositive over several interior values of c . Under recorded-action-quarter estimation, it is small relative to observed event-quarter tax variation, producing very imprecise estimates. For France alone, the fitted endpoint denominator is nonpositive under recorded-action-quarter estimation. I therefore treat these results as supporting diagnostics rather than as robustness evidence comparable to the main U.S. and pooled VAT applications.