Contamination, Bias, and Correction in Text-Based Measurement

Nahyeon Bak* Hanbat Jeong† Daisoon Kim‡

April 2026

Abstract

Natural language processing, machine learning, and artificial intelligence are increasingly used to construct proxies for latent variables from unstructured data in empirical research. We formalize a text-generation environment in which semantic-field activation and lexical co-selection cause word choices to co-vary with observable economic fundamentals, generating state-dependent measurement error in text-based proxies. Thus, using these variables as regressors generates both attenuation bias and contamination-induced omitted-variable bias. Exploiting residual nonlinear dependence, we propose two complementary correction methods: a 2SLS estimator using residualized high-order exogenous variables (RHEV) and a residualized copula-based control-function (RCCF) estimator. Both recover latent-variable effects using only the observed proxy and observable states. Finally, we provide diagnostic tools and practical guidance for empirical applications using text-derived variables. More broadly, our framework highlights that inference with machine-generated variables depends not only on extraction quality but also on the statistical properties of the underlying data-generating process.

JEL Codes: C1; C26; C5; D8.
Keywords: Measurement Error; Text-Based Analysis; Natural Language Processing; Artificial Intelligence; News-Based Uncertainty Index.

*Amazon.com. E-mail: nahyeonbak@gmail.com
†Macquarie University. E-mail: hanbat.jeong@mq.edu.au
‡Andersen Institute of Finance & Economics; NC State University. E-mail: dkim29@ncsu.edu.

The views expressed within are those of the authors and not necessarily those of the Andersen Institute of Finance and Economics, Amazon.com, or their subsidiaries.

1 Introduction

Researchers increasingly rely on natural language processing (NLP) and textual analysis to measure economic variables that are difficult to observe directly. News articles, corporate disclosures, earnings calls, policy documents, and social media posts are routinely transformed into quantitative proxies for uncertainty, risk, sentiment, expectations, political preferences, and other concepts central to the social sciences and business research. Yet despite their widespread adoption, relatively little is known about the econometric properties of these machine-generated measures, particularly how language generation shapes the variables ultimately used for empirical inference.

We study a form of inherent measurement distortion that arises during the generation of text, before any data extraction or econometric analysis takes place. Drawing on semantic-field theory in linguistics (Trier 1931; Lyons 1977), we exploit the insight that word choice is context-dependent. Writers do not select words independently from a flat dictionary. Instead, salient situations activate structured lexical fields whose semantically related expressions become jointly more likely to appear. As emphasized by Wierzbicka (1996) and Murphy (2003), the semantic field activated by a writer depends on the context and on how events are perceived. Consequently, textual descriptions reflect not only the latent concept that a researcher seeks to measure but also broader economic states and conditions that shape language use.

This mechanism creates what we call *textual contamination*. A text-based measure is jointly shaped by the latent variable of interest and by the surrounding economic environment through the language-generation process. As a result, nonclassical measurement error is embedded in text-based proxies: both the proxy and its measurement error may be correlated with economic conditions. Text-based measures therefore inherit information from the environment in which language is produced, not merely from the latent concept they are intended to measure.[1]

News-based uncertainty measures (e.g., Baker, Bloom and Davis 2016) provide a natural illustration. Even holding latent uncertainty fixed, adverse economic conditions make semantic fields containing expressions such as risk, concern, fear, threat, and uncertainty

[1]Notably, our mechanism does not require writers to choose words carelessly or unconsciously. Indeed, careless or random word choice would tend to introduce idiosyncratic measurement error rather than systematic contamination. What matters for our mechanism is that lexical choices respond systematically to the economic environment. Most economic documents are carefully drafted, edited, and subject to substantial institutional or editorial control. Such processes may reduce idiosyncratic variation in language, but they do not imply that lexical choices are invariant to the economic states being described. On the contrary, deliberate word choice may itself respond systematically to economic conditions. Thus, the relevant distinction is not between careful and careless writing, but between state-dependent and state-independent language generation.

more likely to be activated, increasing the prevalence of these words in news coverage during downturns. Because these terms form the basis of many uncertainty dictionaries, the resulting indices reflect not only latent uncertainty but also contemporaneous economic fundamentals. Consequently, news-based uncertainty measures may conflate genuine second-moment uncertainty with first-moment downside conditions.

More broadly, the core insight of this paper extends beyond text-based measurement to the growing use of machine learning (ML) and artificial intelligence (AI) for extracting structured variables from unstructured data. Advances in ML and AI have substantially improved representation and prediction, but they do not alter the underlying data-generating process. If unstructured data systematically encode both latent objects of interest and observable environments, increasingly accurate algorithms will recover both components rather than isolate the latent variable alone. Consequently, credible empirical inference requires accounting for two distinct sources of measurement error: algorithmic measurement/prediction error arising from imperfect statistical learning, and systematic contamination originating from the generation of the underlying unstructured data.

The econometric implications are immediate. When a regression model uses a contaminated proxy as an independent variable, conventional OLS estimation suffers from two distinct distortions. First, measurement error generates the familiar attenuation bias. Second, contamination creates omitted-variable bias because the measurement error itself is correlated with economic states. Depending on the direction of contamination, estimated effects may be biased upward or downward and can even become spuriously significant. More broadly, empirical relationships attributed to latent concepts such as uncertainty, sentiment, political risk, or economic narratives may partly reflect the influence of other economic conditions embedded in the language used to construct those measures, rather than the latent concepts themselves.

A natural response to this bias is to control for related economic state variables directly or to residualize the text-based measure with respect to those variables before estimation. Indeed, an empirical literature using news-based uncertainty measures adopts precisely this strategy by conditioning on first-moment economic variables (e.g., Kim and Lee 2026). We show that such procedures remove part of the contamination-induced bias but generally fail to eliminate attenuation bias. Controlled OLS therefore remains inconsistent. At the same time, an important insight emerges: the same dependence between text-based measures and economic states that generates contamination also contains information useful for identification. In particular, we show that the residual attenuation bias remaining after residualization declines with the strength of contamination itself.[2]

[2]This gain from contamination comes at a cost: stronger contamination increases estimator variance.

Our central identification insight is that removing linear contamination does not eliminate statistical dependence between latent variables and other economic states when they are jointly distributed. Researchers are often interested in estimating the effect of a "pure" latent variable after removing the direct influence of other economic conditions. For example, one may wish to estimate the effect of uncertainty uncorrelated with first-moment economic conditions. Such a goal requires linear orthogonality, not statistical independence. Even when the latent uncertainty is linearly uncorrelated with first-moment fundamentals, non-linear dependence may remain. We show that this residual dependence can be exploited to recover information about the latent variable and to correct the attenuation bias induced by textual contamination. In this sense, our framework transforms a source of contamination into a source of identification.

To operationalize this insight, we develop two complementary correction strategies. The first constructs an instrument variable (IV) from nonlinear transformations of relevant state variables and residualizes them with respect to their linear component, thereby removing their linear alignment with the contamination channel. After removing the component of the text-based measure that is systematically predictable from observed relevant economic state variables, we show that suitably residualized higher-order exogenous variables (RHEVs) can retain information about the latent variable while remaining orthogonal to the contamination channel. Under the corresponding relevance and exclusion conditions, these RHEVs yield a valid two-stage least squares (2SLS) estimator that eliminates the residual attenuation bias left by controlled OLS. Our second approach models the dependence structure between text-based measures and states directly in copula space. Implementing a residualized copula-based control-function (RCCF), we estimate and remove nonlinear dependence that conventional residualization cannot capture. This approach does not rely on instrument relevance and therefore provides a complementary correction strategy when the residualized first stage of the former approach is weak.

The two approaches exploit the same underlying insight but differ in their identifying variation and statistical properties. RHEV-IVs perform particularly well when residual nonlinear dependence provides sufficiently strong relevance, whereas the alternative control-function approach (RCCF) remains informative in weaker-relevance environments. We characterize these trade-offs theoretically and provide practical diagnostic tools to guide empirical implementation.

Our contribution is both conceptual and econometric. Conceptually, we show that text-based measures possess a distinctive form of state-dependent measurement error that arises from the language-generation process itself. Econometrically, we develop a framework for identifying and correcting the resulting bias using only the observed

text-based proxy and economic states, without requiring external instruments, validation datasets, repeated measurements, or auxiliary data sources. Our results suggest that the construction of text-based measures should be viewed not as a passive measurement exercise but as an econometric problem in which language generation, contamination, and identification are fundamentally intertwined.

This paper contributes to the growing literature that uses textual analysis and unstructured data to construct measures of latent economic variables. A seminal contribution is Baker et al. (2016), who develop a news-based measure of economic policy uncertainty and demonstrate its importance for macroeconomic fluctuations. Subsequent research has extended this approach to a wide range of concepts, including financial uncertainty (Alfaro, Bloom and Lin 2024), trade policy uncertainty (Caldara, Iacoviello, Molligo, Prestipino and Raffo 2020), monetary policy uncertainty (Husted, Rogers and Sun 2020), and geopolitical risk (Caldara and Iacoviello 2022). More broadly, advances in NLP, ML, and AI have enabled researchers to extract information from text, images, and audio to study product differentiation (Hoberg and Phillips 2016), asset pricing (Calomiris and Mamaysky 2019), financial regulation (Calomiris, Mamaysky and Yang 2020), monetary policy communication (Gorodnichenko, Pham and Talavera 2023), overseas buyer information (Kim and Kim 2026), and many other economic phenomena. Our paper complements this literature by studying the statistical properties of the resulting machine-generated measures and the implications of language generation for identification and inference.

Our paper is particularly related to the emerging literature on econometric inference with variables constructed from unstructured data. Most closely, Battaglia, Christensen, Hansen and Sacher (2025) show that AI- and ML-generated regressors can produce biased inference because latent variables must first be estimated from text, giving rise to algorithmic measurement error through classification mistakes, topic-model estimation, finite document length, and other first-stage prediction errors. Their focus is on the statistical properties of the extraction algorithm. By contrast, we study a distinct source of measurement error that arises earlier, during the generation of the underlying text. We show that context-dependent semantic-field activation causes language to encode both the latent concept of interest and observable environments.

Recent work has developed econometric methods for correcting algorithmic measurement error in AI- and ML-generated regressors (e.g., Yang, Adomavicius, Burtch and Ren 2018; Yang, McFowland, Burtch and Adomavicius 2022; Burtch, McFowland, Yang and Adomavicius 2026; Wei and Malik 2026). A defining feature of this literature is that supervised learning provides labeled training or validation samples, making prediction error at least partially observable and therefore amenable to conventional measurement-

error corrections.[3] Our setting differs fundamentally because the source of bias is not the extraction algorithm but the underlying unstructured data themselves. Since language is generated through context-dependent semantic-field activation, the resulting corpus systematically reflects both latent concepts and observable economic states. Improvements in NLP, ML, or AI algorithms can therefore reduce prediction/representation error, but they do not necessarily eliminate—and may even recover more precisely—contamination arising from the underlying data-generating process.

The remainder of the paper proceeds as follows. Section 2 introduces the text-generation framework to show how state-dependent contamination arises in text-based measures. Section 3 characterizes the econometric consequences of this contamination in conventional regressions. Section 4 develops two complementary correction methods and establishes their properties, inference, and diagnostic procedures. Section 5 evaluates the finite-sample performance of the proposed estimators. Section 6 provides guidance for selecting between the two correction methods. The final section concludes.

2 Text-Based Measures as Contaminated Proxies

Textual analysis transforms unstructured language into quantitative proxies for otherwise unobservable economic concepts. Their usefulness, however, depends on how language is generated. Texts are not produced by observing a latent variable of interest in isolation and then describing it mechanically. Rather, writers describe economic events using narratives, semantic fields, and lexical choices that respond jointly to the latent concept of interest and the broader economic environment. Consequently, text-based measures can systematically inherit exposure to observable economic states even when the research objective is to measure a latent variable purged of those states.

News-based measures of uncertainty provide a useful illustration with a clear direction of contamination. During adverse economic conditions, journalists are more likely to describe events using lexical fields containing words such as risk, uncertainty, or volatility. As a result, uncertainty indices constructed from these words may capture not only latent second-moment uncertainty but also contemporaneous deterioration in economic conditions and fundamentals. This section develops a general framework for formalizing this mechanism.

The first subsection shows that a wide range of text-based measures can be reduced

[3] Supervised learning relies on manually labeled training and validation data, so the true target variable is observed for a subset of observations. This allows researchers to estimate the prediction-error distribution and apply measurement-error correction methods.

to a single conditional probability. We then explain why this probability depends on the surrounding economic environment (state-dependent) through semantic-field activation and lexical co-selection. Finally, we introduce a reduced-form representation of text-based measurement and contamination to derive the econometric implications.

2.1 Text-Generation and Text-Based Measurement

Environment For each period t, let (x_t, w_t) denote the economic state, where x_t is the latent variable of interest and w_t captures observable economic states. The researcher's object of interest is the latent variable x_t, interpreted as the *pure* component of the underlying concept after separating it from the direct influence of other states. For example, in the context of uncertainty measurement, x_t represents pure second-moment uncertainty, whereas w_t captures observable first-moment macroeconomic conditions and fundamentals.

The latent variable x_t is unobserved. Instead, researchers observe (x_t^*, w_t), where x_t^* is a proxy constructed from an underlying text corpus. The corpus for period t consists of N_t documents. Document d contains $L_{d,t}$ tokens, and $\tau_{l,d,t} \in \mathcal{V}$ denotes the l-th token, where $\mathcal{V}$ is the token vocabulary. Let $\mathcal{V}^* \subseteq \mathcal{V}$ denote the collection of words and expressions associated with the latent concept of interest. For example, when measuring uncertainty, $\mathcal{V}^*$ contains uncertainty-related words and phrases such as "uncertainty," "risk," and "volatility."[4]

Statistical Representation of Text-Based Measures Most text-based measures used in economics can be viewed as functions of the occurrence of vocabulary in $\mathcal{V}^*$. Although the construction varies across applications, two classes of measures are particularly common. We first consider word-frequency measures. The proxy is defined as the average share of target vocabulary appearing in the corpus,

$$x_t^* \equiv \frac{1}{N_t} \sum_{d=1}^{N_t} \frac{1}{L_{d,t}} \sum_{l=1}^{L_{d,t}} \mathbb{I}\{\tau_{l,d,t} \in \mathcal{V}^*\},$$

[4]Strictly speaking, words and tokens are distinct units of textual analysis. A token represents a subword, whole word, punctuation mark, or character segment generated via algorithms such as byte-pair encoding to transform raw text into discrete numerical indices for transformer-based architectures. Empirically, a single token corresponds to approximately three to four characters, or roughly 0.75 words in standard English text. While tokenization granularity can influence downstream representations, our analysis abstracts from tokenization-specific mechanics; consequently, we use the terms "words" and "tokens"—as well as "vocabulary" and "token vocabulary"—interchangeably hereafter.

where $\mathbb{I}\{\cdot\}$ is the indicator function. Conditional on the state, (x_t, w_t), its mean is

$$\mathbb{E}[x_t^*|x_t, w_t] = \Pr(\tau_{l,d,t} \in \mathcal{V}^*|x_t, w_t). \tag{1}$$

Thus, apart from sampling variability, the proxy measures the probability that vocabulary from the target lexical field is generated.

We next consider article-frequency measures. Here the proxy equals the fraction of documents containing at least one occurrence of vocabulary in $\mathcal{V}^*$,

$$x_t^* \equiv \frac{1}{N_t} \sum_{d=1}^{N_t} \left(1 - \prod_{l=1}^{L_{d,t}} \mathbb{I}\{\tau_{l,d,t} \notin \mathcal{V}^*\}\right).$$

Conditional on (x_t, w_t), tokens are drawn independently across positions within a document, so the probability that a document of length $L_{d,t}$ contains no target vocabulary is $\big[1 - \Pr(\tau_{l,d,t} \in \mathcal{V}^*|x_t, w_t)\big]^{L_{d,t}}$. Taking the complement and averaging across documents, the conditional mean is

$$\mathbb{E}[x_t^*|x_t, w_t] = \frac{1}{N_t} \sum_{d=1}^{N_t} \left\{1 - \big[1 - \Pr(\tau_{l,d,t} \in \mathcal{V}^*|x_t, w_t)\big]^{L_{d,t}}\right\}, \tag{2}$$

which is an increasing function of the same conditional probability that a target word is generated.[5]

The two measures therefore share a common statistical structure. Although one aggregates word frequencies and the other document frequencies, both depend on the probability that vocabulary associated with the latent concept is generated conditional on the underlying state. The following lemma establishes a common representation.

Lemma 1 *(Representation of Text-Based Measures) Conditional on (x_t, w_t), suppose tokens are drawn independently across positions and document length, $L_{d,t}$, is a measurable function of the state, (x_t, w_t). Then,*

$$\mathbb{E}[x_t^*|x_t, w_t] \equiv G\big(\Pr(\tau_{l,d,t} \in \mathcal{V}^*|x_t, w_t)\big),$$

where $G(\cdot)$ is a monotone increasing function.

[5]The mapping from $\Pr(\tau_{l,d,t} \in \mathcal{V}^*|x_t, w_t)$ to $\mathbb{E}[x_t^*|x_t, w_t]$ in (2) involves the document lengths $\{L_{d,t}\}_{d=1}^{N_t}$ through the exponents. To express the conditional mean as a function of the scalar $\Pr(\tau_{l,d,t} \in \mathcal{V}^*|x_t, w_t)$ alone, we treat document length as a measurable function of the state (x_t, w_t). Since each term $1 - \big[1 - \Pr(\tau_{l,d,t} \in \mathcal{V}^*|x_t, w_t)\big]^{L_{d,t}}$ is increasing in the underlying probability, so is their average, and $G(\cdot)$ is well defined as a function of that probability. The constant-length case $L_{d,t} = L_t$, for which $\mathbb{E}[x_t^*|x_t, w_t] = 1 - \big[1 - \Pr(\tau_{l,d,t} \in \mathcal{V}^*|x_t, w_t)\big]^{L_t}$, is a special case.

Lemma 1 shows that a broad class of text-based measures can be viewed as monotone transformations of a common underlying object: the conditional probability that vocabulary from the target lexical field is generated. Consequently, understanding the statistical properties of text-based measures reduces to understanding the determinants of $\Pr(\tau_{l,d,t} \in \mathcal{V}^* | x_t, w_t)$. If vocabulary generation is driven solely by the latent variable, text-based measures behave like conventional noisy proxies. If vocabulary generation also depends on the other surrounding economic environments, however, the resulting proxy inherits systematic exposure to those environments. The next subsection characterizes this state-dependence by developing a text-generation model based on semantic-field activation and lexical co-selection.

2.2 Writing-Stage Contamination

The previous subsection showed that the statistical properties of text-based measures are governed by the conditional probability, $\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid x_t, w_t)$. This subsection develops a language-generation framework for this probability. Our central point is that contamination arises before the econometrician constructs a text-based index. It originates during the writing process itself. Because writers describe economic events using semantic fields rather than isolated words, lexical choices respond jointly to the latent concept of interest and to the surrounding economic environment. Consequently, text-based measures may inherit systematic exposure to observable states even before any dictionary, classifier, topic model, or large language model is applied.

Semantic-Field Theory Our framework builds on semantic-field theory in linguistics. Beginning with Trier (1931) and further developed by Lyons (1977), semantic-field theory views vocabulary as organized into interconnected lexical fields rather than as isolated words. When a particular situation becomes salient, writers activate an entire semantic field whose members are semantically related and therefore tend to be selected together. Murphy (2003) studies these lexical relations in detail, while Wierzbicka (1996) emphasizes that semantic-field activation depends systematically on context.

In the present setting, this means that a journalist describing an economic state is not simply choosing whether to include a specific word; the writer is activating a larger field of mutually reinforcing expressions that fit the perceived character of the state being described. More specifically, when writers face or cover bad economic conditions, an activated semantic field is different from that of good economic conditions. In other words, the semantic field activated at the writing stage is connected to the broader economic

states that writers face or that characterize the situations the articles cover. This semantic field contains words and phrases that mechanically overlap with the vocabulary used to construct indices.

Text-Generation Framework Let $\mathcal{F}_t \in \mathbb{F}$ denote the semantic field activated during the generation of the relevant text in period t, and let $\mathbb{F}$ denote the collection of possible semantic fields. Throughout, we hold the latent variable of interest x_t fixed, so that all statements below isolate the effect of the surrounding economic states on vocabulary generation at a given level of the latent state.

Our first assumption formalizes the central implication of semantic-field theory (Trier 1931; Lyons 1977; Murphy 2003): lexical co-selection.

Assumption 1 (Lexical Overlap) *Different semantic fields exhibit different probabilities of generating vocabulary belonging to the target lexical field:*

$$\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}_t = \mathcal{F}, x_t) \neq \Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}_t = \mathcal{F}', x_t), \qquad \mathcal{F} \neq \mathcal{F}'.$$

Assumption 1 simply states that different semantic fields employ different vocabularies. Consequently, the probability that words from the target lexical field appear varies across activated semantic fields.

The second assumption characterizes semantic-field activation.

Assumption 2 (Semantic-Field Activation) *The distribution over semantic fields depends on the observable economic state, holding the latent state fixed:*

$$\Pr(\mathcal{F}_t = \mathcal{F} \mid x_t, w_t = w) \neq \Pr(\mathcal{F}_t = \mathcal{F} \mid x_t, w_t = w'), \qquad w \neq w'.$$

Assumption 2 states that the probability of activating a semantic field varies systematically with economic conditions, even after conditioning on the latent variable of interest. It imposes no restrictions on the mechanism generating this dependence. Rather, it requires only that the surrounding environment shifts the distribution over semantic fields.

Several complementary mechanisms can generate such shifts. On the demand side, changes in economic conditions may alter the value of information about particular topics, thereby changing demand for related narratives (e.g., Van Nieuwerburgh and Veldkamp 2006; Fajgelbaum, Schaal and Taschereau-Dumouchel 2017). On the behavioral side, cognitive asymmetries such as loss aversion, selective attention, and negativity bias imply that different environments make different concepts more cognitively salient (e.g., Kahneman and Tversky 1979; Baumeister, Bratslavsky, Finkenauer and Vohs 2001; Rozin and

Royzman 2001). On the supply side, writers, journalists, or other content producers may systematically emphasize different narratives in response to changing economic conditions or audience demand (e.g., Soroka 2006; Soroka, Daku, Hiaeshutter-Rice, Guggenheim and Pasek 2018; Damstra and Boukes 2021). Although these mechanisms differ in their economic, psychological, and institutional foundations, they share the common implication that the distribution of activated semantic fields depends on observable states.

Our framework takes a neutral stance regarding which of these mechanisms works in practice. The contamination result in Lemma 2 requires only Assumption 2; it does not depend on the particular source of semantic-field activation. The direction of contamination is thus application-specific and is determined jointly by (i) how broader economic conditions shift the distribution over semantic fields and (ii) how those fields overlap with the vocabulary used to construct the text-based proxy. In the leading application of news-based uncertainty measures, adverse economic conditions increase the probability of activating downside semantic fields, which overlap disproportionately with uncertainty-related vocabulary. In Section 2.3, this directional restriction is explicitly imposed.

Probability of Generating Vocabulary Applying the law of total probability yields

$$\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid x_t, w_t) = \sum_{\mathcal{F} \in \mathbb{F}} \Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}_t = \mathcal{F}, x_t, w_t) \Pr(\mathcal{F}_t = \mathcal{F} \mid x_t, w_t). \tag{3}$$

To isolate the semantic-field channel, we impose one additional restriction.

Assumption 3 (Semantic-Field Mediation) *Conditional on the activated semantic field and the latent state, the other observable economic states do not further affect vocabulary selection:*

$$\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}_t = \mathcal{F}, x_t, w_t) = \Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}_t = \mathcal{F}, x_t).$$

This additional assumption isolates contamination arising through semantic-field activation alone. Allowing the surrounding economic conditions to influence lexical choice even within a given semantic field would introduce an additional direct channel and would strengthen, rather than weaken, our results.

Substituting Assumption 3 into equation (3) yields

$$\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid x_t, w_t) = \sum_{\mathcal{F} \in \mathbb{F}} \Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}_t = \mathcal{F}, x_t) \Pr(\mathcal{F}_t = \mathcal{F} \mid x_t, w_t). \tag{4}$$

Then, the following lemma summarizes the implication.

Lemma 2 (Writing-Stage Contamination) *Suppose Assumptions 1–3 hold. Then, holding the latent state x_t fixed,*

$$\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid x_t, w_t = w) \neq \Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid x_t, w_t = w'), \qquad w \neq w',$$

whenever economic conditions affect semantic-field activation.

This lemma establishes that, even at a fixed level of the latent variable, the probability of generating vocabulary from the target lexical field systematically varies with the surrounding economic environment. Combining this result with Lemma 1 implies that the resulting text-based proxy inherits systematic exposure to economic conditions before any econometric analysis is performed.

Measurement Error of Text-Based Proxy The writing-stage mechanism admits a convenient probabilistic interpretation. Applying Bayes' rule to the conditional probability that vocabulary from the target lexical field is generated yields

$$\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid x_t, w_t) = \underbrace{\frac{\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid w_t)}{\Pr(\tau_{l,d,t} \in \mathcal{V}^*)}}_{\substack{\text{economic states-driven} \\ \text{vocabulary tilt}}} \times \underbrace{\Pr(\tau_{l,d,t} \in \mathcal{V}^*) \frac{\Pr(x_t \mid \tau_{l,d,t} \in \mathcal{V}^*, w_t)}{\Pr(x_t \mid w_t)}}_{\substack{\text{informativeness for} \\ \text{the latent state}}}, \tag{5}$$

where the second component reflects how informative the selected vocabulary is about the latent variable of interest.

The first component isolates the role of the surrounding environment emphasized in Lemma 2. It compares the probability that vocabulary from the target lexical field is generated under the surrounding economic environment with its unconditional probability. When the ratio equals one, economic conditions do not distort vocabulary choice and the proxy behaves as a conventional text-based measurement. Deviations from one imply that semantic-field activation systematically tilts the vocabulary distribution. This representation is a heuristic that makes the source of contamination transparent; it is not the object we ultimately estimate, and the formal results below do not rely on it.

The state-dependent nature of the resulting measurement error follows directly from the two preceding propositions rather than from this Bayes-rule interpretation. Lemma 2 establishes that the conditional probability, $\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid x_t, w_t)$, varies with observable economic environments at a fixed latent state, and Lemma 1 maps this probability monotonically into the conditional mean of the proxy. Combining the two yields the following corollary.

Proposition 1 (State-Dependent Measurement Error) *Suppose the conditions of Lemmas 1 and 2 hold. Then*

$$\mathbb{E}[x_t^* \mid x_t, w_t] \neq \mathbb{E}[x_t^* \mid x_t].$$

Consequently, the text-based proxy is not subject to classical white-noise measurement error alone. Instead, the measurement error is nonclassical: the conditional distribution of both the proxy and its measurement error depends on observable economic states through the language-generation process.

2.3 Example: News-Based Uncertainty Index

News-based uncertainty indices provide a natural illustration of the writing-stage contamination mechanism developed above. The latent object of interest is pure uncertainty—a second-moment concept reflecting the dispersion of future outcomes. Economic fundamentals, by contrast, capture first-moment prevailing economic conditions such as current output or employment. Our framework predicts that these two objects become statistically intertwined because adverse economic conditions activate semantic fields that overlap disproportionately with uncertainty-related vocabulary. Consequently, a news-based uncertainty index may increase even when latent uncertainty itself is unchanged.

For expositional purposes, suppose there are two stylized semantic fields corresponding to downside and upside economic narratives: $\mathbb{F} = \{\mathcal{F}_\ominus, \mathcal{F}_\oplus\}$.

$$\mathcal{F}_\ominus = \{\text{uncertain, volatile, risk, vulnerable, exposure, fragile, downside, } \ldots\}.$$

By contrast, upside states are more likely to activate a different lexical field, such as

$$\mathcal{F}_\oplus = \{\text{opportunity, prospect, growth, expansion, recovery, gains, } \ldots\}.$$

Let $\mathcal{V}^*$ denote the uncertainty dictionary used to construct the index. Consistent with Assumption 1, uncertainty-related vocabulary is more likely to be selected within the downside semantic field than within the upside semantic field:

$$\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}_\ominus, x_t) > \Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}_\oplus, x_t). \tag{6}$$

Moreover, adverse economic fundamentals increase the probability that the downside semantic field is activated. Consistent with Assumption 2,

$$\Pr(\mathcal{F}_\ominus \mid x_t, w) > \Pr(\mathcal{F}_\ominus \mid x_t, w'), \qquad w < w', \tag{7}$$

where larger values of w_t correspond to stronger economic conditions (e.g., expansion $= w' > w =$recession). This directional restriction specializes Assumption 2 to the case of news-based uncertainty measures. It is consistent with evidence that downturns increase the demand for uncertainty-related information (e.g., Van Nieuwerburgh and Veldkamp 2006), that losses and adverse outcomes receive disproportionate cognitive attention (e.g., Kahneman and Tversky 1979; Baumeister et al. 2001; Rozin and Royzman 2001), and that the media disproportionately frame adverse economic conditions through downside narratives (e.g., Soroka 2006; Soroka et al. 2018; Damstra and Boukes 2021). Combined with the stronger lexical overlap in equation (6), these mechanisms imply that uncertainty-related vocabulary becomes systematically more prevalent during economic downturns, even after conditioning on the latent uncertainty state.

To isolate the semantic-field channel, we maintain the mediation (Assumption 3):

$$\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}, x_t, w_t) = \Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}, x_t). \tag{8}$$

Then, the following corollary specializes Lemma 2 and Proposition 1 to the case of news-based uncertainty measures with the clear direction of contamination.

Corollary 1 (Downside Contamination of News-Based Uncertainty Index) *Suppose equations (6)–(8) hold. Then, for $w < w'$,*

$$\begin{aligned}&\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid x_t, w) - \Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid x_t, w') \\ &= \left[\Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}_\ominus, x_t) - \Pr(\tau_{l,d,t} \in \mathcal{V}^* \mid \mathcal{F}_\oplus, x_t)\right] \times \left[\Pr(\mathcal{F}_\ominus \mid x_t, w) - \Pr(\mathcal{F}_\ominus \mid x_t, w')\right] > 0\end{aligned}$$

Consequently,

$$\mathbb{E}[x_t^* \mid x_t, w] > \mathbb{E}[x_t^* \mid x_t, w'], \qquad w < w',$$

even when the latent uncertainty state, x_t, is unchanged.

This corollary isolates the mechanism of downside contamination in news-based uncertainty measures through which writing-stage contamination arises. Consider two economies with identical latent uncertainty but different macroeconomic conditions. The economy experiencing weaker fundamentals activates downside semantic fields more frequently, making uncertainty-related vocabulary more likely to appear in otherwise comparable news articles. Consequently, the observed uncertainty index rises despite no change in the underlying second-moment uncertainty.

Discussion When Baker et al. (2016) construct their newspaper-based uncertainty index, they not only count newspaper articles containing economy, policy, and uncertainty

terms, but also validate the resulting measure through extensive human audits. The audit procedure establishes that human readers consistently identify articles intended to communicate uncertainty. From the perspective of our framework, this validates the semantic interpretation of the constructed index. At the same time, because both journalists and auditors rely on the same context-dependent semantic cues, the audit cannot distinguish latent uncertainty from writing-stage contamination induced by economic fundamentals.

Our objective is not to replace existing news-based uncertainty measures or question their empirical usefulness. Rather, we study the econometric implications of using such measures as independent variables. We argue that observed news-based uncertainty indices generally combine information about latent uncertainty with information about prevailing economic conditions. More broadly, whenever text is generated through context-dependent semantic-field activation, observed text-based proxies need not isolate the latent variable of interest. The next subsection abstracts from this application and develops a reduced-form characterization of writing-stage contamination that applies to a broad class of text-based measures. This representation forms the basis for the identification, estimation, and inference procedures developed in the remainder of the paper.

2.4 Reduced-Form Contamination

Economic states affect semantic-field activation and lexical co-selection during the text-generation process, thereby changing the probability that target vocabulary is used before any econometric measurement takes place. While the linguistic mechanism provides the micro-foundation for writing-stage contamination, it is neither necessary nor desirable to model every stage of language generation in subsequent econometric analysis. Instead, we summarize these mechanisms through a reduced-form measurement equation that captures their implications for the observed proxy. The resulting representation applies broadly to econometric analysis with text-based measures and, more generally, to variables constructed from unstructured data such as text, images, audio, or video. Figure 1 summarizes the mechanism developed in the previous subsections with the reduced-form representation with linearized contamination of this subsection.

Let $g(x, w)$ denote the conditional measurement mapping induced by the text-generation process:

$$g(x, w) \equiv \mathbb{E}[x_t^* \mid x_t = x, w_t = w]. \tag{9}$$

Under classical measurement error, the expected proxy depends only on the latent variable,

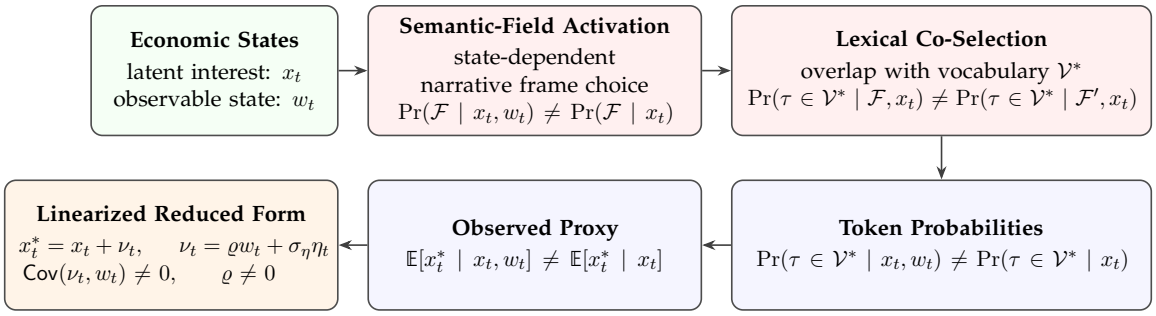


Figure 1: Writing-Stage Contamination in Text-Based Measures.

so that $g(x, w) = g(x)$. Writing-stage contamination instead implies

$$\frac{\partial g(x, w)}{\partial w} \neq 0,$$

meaning that observable economic states systematically shift the expected value of the proxy even after conditioning on the latent variable of interest.

Throughout the econometric analysis, we work with a first-order approximation of the conditional measurement mapping around a reference state. After recentering the variables and normalizing the scale of the latent variable, the mapping admits the local approximation:

$$g(x_t, w_t) \approx x_t + \varrho w_t, \qquad \text{where } \varrho \equiv \left.\frac{\partial g(x, w)}{\partial w}\right|_{(0,0)}. \tag{10}$$

The coefficient ϱ measures the sensitivity of the proxy to economic conditions. This linearization immediately yields the following reduced-form representation.[6]

Proposition 2 (Reduced-Form Representation: Linear Contamination Model) *Under the local approximation of the conditional measurement mapping, the text-based proxy admits the representation:*

$$x_t^* = x_t + \nu_t, \qquad \nu_t = \varrho w_t + \sigma_\eta \eta_t, \tag{11}$$

where η_t is an idiosyncratic measurement component satisfying $\mathbb{E}[\eta_t \mid w_t] = 0$ and $\mathsf{Var}(\eta_t) = 1$, and σ_η governs the magnitude of this idiosyncratic noise.

The reduced-form decomposition separates two distinct sources of measurement error.

[6]The linear representation in equation (11) is a local approximation to the measurement mapping. More generally, one may write $x_t^* = x_t + \varrho w_t + q(w_t)'\delta + \sigma_\eta \eta_t$, where $q(w_t)$ collects nonlinear transformations of the observable economic state variable. The appendix develops this extension and the corresponding orthogonalization strategy. The main text focuses on the linear component because it delivers the central sign implication most transparently.

The first component, ϱw_t, is systematic and arises from writing-stage contamination. It reflects the dependence of language generation on the surrounding economic environment established in the previous subsections. The second component, $\sigma_\eta \eta_t$, captures conventional idiosyncratic measurement error arising from finite corpus size, finite document length, imperfect dictionary design, classification error, or other implementation-specific features of text measurement. Unlike writing-stage contamination, this component is unrelated to contemporaneous economic conditions.

For news-based uncertainty measures, larger values of w_t correspond to stronger macroeconomic conditions. The downside contamination mechanism developed in Section 2.3 therefore implies $\varrho < 0$, so that $\mathsf{Cov}(\nu_t, w_t) < 0$. The measurement error is therefore state-dependent rather than classical: adverse economic conditions systematically increase measured uncertainty even when latent uncertainty itself is unchanged.

3 Traditional Regressions with a Text-Based Proxy

This section studies the econometric consequences of writing-stage contamination. We first formulate the regression problem faced by empirical researchers who observe only the contaminated proxy. We then characterize the probability limits of conventional OLS estimators and show that neither controlling for observable economic state variables nor residualizing the proxy eliminates attenuation arising from idiosyncratic text-measurement noise. These results motivate the correction procedures developed in the next section.

3.1 Text-Based Proxy Regressor

Throughout the remainder of the paper, we study the consequences of using a contaminated text-based proxy in a linear regression. Consider the structural outcome equation

$$y_t = \beta x_t + \gamma w_t + \varepsilon_t, \tag{12}$$

where all variables are treated as centered for simplicity. The coefficient β measures the *pure* effect of the latent variable (x_t) with $\mathsf{Cov}(x_t, w_t) = 0$ holding the relevant economic state level (w_t) fixed. Recovering this parameter is the objective of researchers.

Since x_t is latent, empirical researchers replace it with the observed text-based proxy, x_t^*. Combining the structural equation with the reduced-form contamination model in Proposition 2 yields

$$y_t = \beta x_t^* + \gamma w_t + u_t, \qquad u_t = \varepsilon_t - \beta \nu_t. \tag{13}$$

Because the measurement error ν_t is embedded in the observed proxy,

$$\mathsf{Cov}(x_t^*, u_t) = -\beta \mathsf{Var}(\nu_t) \neq 0,$$

so that the proxy is endogenous even when the latent variable itself is exogenous.

Equation (13) illustrates that writing-stage contamination generates two distinct econometric problems. The idiosyncratic component of ν_t produces the familiar attenuation bias associated with classical measurement error. The systematic component, ϱw_t, instead embeds information about observable state variables directly into the proxy, creating an additional source of endogeneity that is absent under classical measurement error. Consequently, regressions using text-based proxies may confound the causal effect of the latent variable with the effect of other economic states incorporated into the proxy during the language-generation process.

This distinction is particularly important because text-based measures are frequently used to test whether a latent concept has a statistically significant effect on economic outcomes. A significant coefficient on the observed proxy therefore need not imply a significant effect of the latent variable itself. Instead, it may simply reflect information about the surrounding economic environment that has been mechanically incorporated into the proxy through writing-stage contamination.

3.2 Traditional OLS Regressions

This subsection characterizes the probability limits of conventional estimators under text-based measurement's contamination and establishes the benchmark against which the proposed correction methods are evaluated. Our analysis builds on the structural equation (12) together with the reduced-form measurement representation in Proposition 2. We begin with imposing the following regularity conditions.

Assumption 4 (Common Regularity and Contamination) *The process $\{y_t, x_t, w_t, \nu_t, \varepsilon_t\}$ is strictly stationary and weakly dependent, with second moments for the probability limits below and fourth moments for the variance estimates. The variables are mean-zero, and*

$$\sigma_x^2 \equiv \mathsf{Var}(x_t) > 0, \qquad \sigma_w^2 \equiv \mathsf{Var}(w_t) > 0.$$

The benchmark orthogonality and contamination conditions are

$$\mathsf{Cov}(x_t, w_t) = 0, \qquad \mathsf{Cov}(x_t, \nu_t) = 0, \qquad \mathsf{Cov}(x_t, \varepsilon_t) = 0, \qquad \mathsf{Cov}(w_t, \varepsilon_t) = 0,$$

$$\mathsf{Cov}(\varepsilon_t, \nu_t) = 0, \qquad \mathsf{Cov}(\eta_t, w_t) = 0, \qquad \mathsf{Cov}(\nu_t, w_t) \neq 0 \quad (\varrho \neq 0).$$

Under this parameterization, the induced second moments of the measurement error are

$$\sigma_\nu^2 \equiv \mathsf{Var}(\nu_t) = \varrho^2 \sigma_w^2 + \sigma_\eta^2, \qquad \sigma_{\nu,w} \equiv \mathsf{Cov}(\nu_t, w_t) = \varrho \sigma_w^2, \qquad \rho_{\nu,w} \equiv \frac{\sigma_{\nu,w}}{\sigma_\nu \sigma_w} = \frac{\varrho \sigma_w}{\sqrt{\varrho^2 \sigma_w^2 + \sigma_\eta^2}}.$$

The normalization, $\mathsf{Cov}(x_t, w_t) = 0$, formalizes the notion that β represents the *pure* effect of our latent object after removing the surrounding economic environment. The condition of $\mathsf{Cov}(x_t, \nu_t) = 0$ isolates the nonclassical feature emphasized here: the measurement error is allowed to be correlated with the observable economic state (contamination, $\mathsf{Cov}(\nu_t, w_t) \neq 0$), but not with the latent uncertainty signal itself.

OLS Regressions We begin by studying two conventional estimators commonly used in empirical work. The first ignores observable states, while the second controls for them directly. This control approach is a natural response to writing-stage contamination.

Definition 1 (OLS Benchmarks) *Let $\widehat{\widetilde{b}}$ denote the coefficient from the naive regression of y_t on x_t^* only. Let $\widehat{b}$ denote the coefficient on x_t^* from the controlled regression of y_t on (x_t^*, w_t).*

An alternative way to remove writing-stage contamination is to residualize the proxy with respect to observable states before estimation. Under linear projection, the Frisch-Waugh-Lovell theorem implies that this procedure is numerically equivalent to including economic state variables as controls.

The population linear projection of the proxy onto the observable state has slope:

$$\frac{\mathsf{Cov}(w_t, x_t^*)}{\mathsf{Var}(w_t)} = \frac{\mathsf{Cov}(w_t, \nu_t)}{\mathsf{Var}(w_t)} = \frac{\sigma_{\nu,w}}{\sigma_w^2} = \varrho,$$

where the second equality follows from the benchmark orthogonality condition, $\mathsf{Cov}(x_t, w_t) = 0$. Consequently, the population residualized proxy is

$$x_t^{*\perp} \equiv x_t^* - \varrho w_t = x_t + \sigma_\eta \eta_t. \tag{14}$$

Residualization therefore removes the systematic component, ϱw_t, of writing-stage contamination, leaving only the idiosyncratic measurement error, $\sigma_\eta \eta_t$.

3.3 Properties of Traditional OLS Estimators

Theorem 1 (Probability Limits of OLS Regressions) *Suppose Assumption 4 holds. Then,*

$$\operatorname{plim} \widehat{\widehat{b}} = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\nu^2} + \frac{\gamma \sigma_{\nu,w}}{\sigma_x^2 + \sigma_\nu^2},$$

and

$$\operatorname{plim} \widehat{b} = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\nu^2(1 - \rho_{\nu,w}^2)} = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\eta^2} = \operatorname{plim} \widehat{b}_\perp.$$

Theorem 1 yields three immediate implications. First, naive OLS is affected by two distinct sources of bias. The first $-\beta \frac{\sigma_\nu^2}{\sigma_x^2+\sigma_\nu^2}$ arises from measurement error, whereas the second $\frac{\gamma \sigma_{\nu,w}}{\sigma_x^2+\sigma_\nu^2}$ reflects the systematic dependence of the proxy on observable states. Interestingly, the contamination ($\rho_{\nu,w}$ equivalently ϱ) affects naive OLS through two distinct channels. Strong contamination directly increases through the covariance term $|\sigma_{\nu,w}|$. At the same time, it increases the overall variance of the proxy error, σ_ν^2, thereby strengthening attenuation. The overall bias therefore reflects the interaction of these two forces.

Second, controlling for observable states removes the component of the measurement error that is linearly aligned with w_t. Consequently, the probability limit depends only on the idiosyncratic text-measurement component, σ_η^2 instead of the total measurement-error variance, $\sigma_\nu^2 = \varrho^2 \sigma_w^2 + \sigma_\eta^2$. Consequently, controlled OLS is an important benchmark but does not consistently recover the structural parameter. Interestingly, the remaining attenuation after controlling for observable states depends on $\sigma_\eta^2 = \sigma_\nu^2(1 - \rho_{\nu,w}^2)$. Thus, stronger contamination mechanically increases the effectiveness of controlling for observable states by removing a larger fraction of total measurement error.

Third, projection-based decontamination, equivalently controlled OLS, completely removes the linear contamination component but does not eliminate attenuation bias. The residualized proxy remains an imperfect measure of the latent variable because it still contains the idiosyncratic text-measurement component $\sigma_\eta \eta_t$. These results show that residualization, or controlling for observable states, is a necessary first step for eliminating systematic contamination, but it is not sufficient for consistent estimation.

Furthermore, controlled/residualized OLS does not necessarily exhibit a smaller overall bias than naive OLS, while it always dominates naive OLS in terms of attenuation.

$$\left| \operatorname{plim} \widehat{\widehat{b}} - \beta \right| < \left| \operatorname{plim} \widehat{b} - \beta \right| \quad \text{if and only if} \quad \left| 1 - \frac{\gamma \rho_{\nu,w}}{\beta} \frac{\sigma_w}{\sigma_\nu} \right| < 1 - \frac{\rho_{\nu,w}^2 \sigma_x^2}{\sigma_x^2 + (1 - \rho_{\nu,w}^2)\sigma_\nu^2},$$

where the right-hand side is not greater than one.

When $\gamma \rho_{\nu,w}/\beta > 0$ (equivalently, $\gamma \varrho/\beta > 0$), it is possible that the naive regression has a

smaller absolute bias precisely. This result should not be interpreted as a justification for naive OLS. Rather, it reflects a cancellation between two distinct misspecification forces: attenuation bias and contamination bias. Because the degree of cancellation depends on unknown population parameters, it cannot be exploited in empirical applications.

The key econometric implication is that writing-stage contamination and classical measurement error are fundamentally different problems. Controlling for observable states or residualizing the proxy removes the systematic contamination component but leaves the idiosyncratic measurement error unchanged. Consequently, conventional OLS procedures cannot consistently recover the structural parameter even after decontamination. This observation motivates the correction methods developed in the next section, which improve residualization/control approach to remove the remaining attenuation bias.

4 Proposed Correction Methods

The previous section shows that controlling for the surrounding environment or using a residualized proxy corrects systematic writing-stage contamination but leaves attenuation bias. This section develops two complementary correction methods that address this remaining attenuation problem by exploiting variation in observable states beyond their linear component. The first method uses higher-order transformations of the observed fundamentals, residualized with respect to their first-order component, as excluded instruments. We refer to these as *residualized higher-order exogenous variables* (RHEVs). The second method instead captures nonlinear dependence between the residualized proxy and observable states on a copula scale through a control function. We refer to this approach as the *residualized copula-based control function* (RCCF). Figure 2 summarizes the relationship between the OLS benchmarks and these two complementary correction strategies.

Shared Source of Identifying Information Both methods draw on the same identifying variation, which we describe before presenting either estimator. Our framework starts from the premise that the latent variable x_t and the observable state variable w_t are generated by the same underlying economic state in period t, and are therefore unlikely to be statistically independent even though the benchmark normalization sets $\mathsf{Cov}(x_t, w_t) = 0$. The key point is that linear uncorrelatedness is a single restriction, whereas statistical independence is infinitely many: $\mathsf{Cov}(x_t, w_t) = 0$ constrains only the first cross-moment $\mathbb{E}[x_t w_t]$, leaving the higher-order cross-moments $\mathbb{E}[x_t^j w_t^k]$ with $j + k \geq 3$—and the joint distribution beyond that single condition—entirely free. The latent signal and the observable economic state can thus be linearly uncorrelated yet strongly dependent through nonlinear and distributional

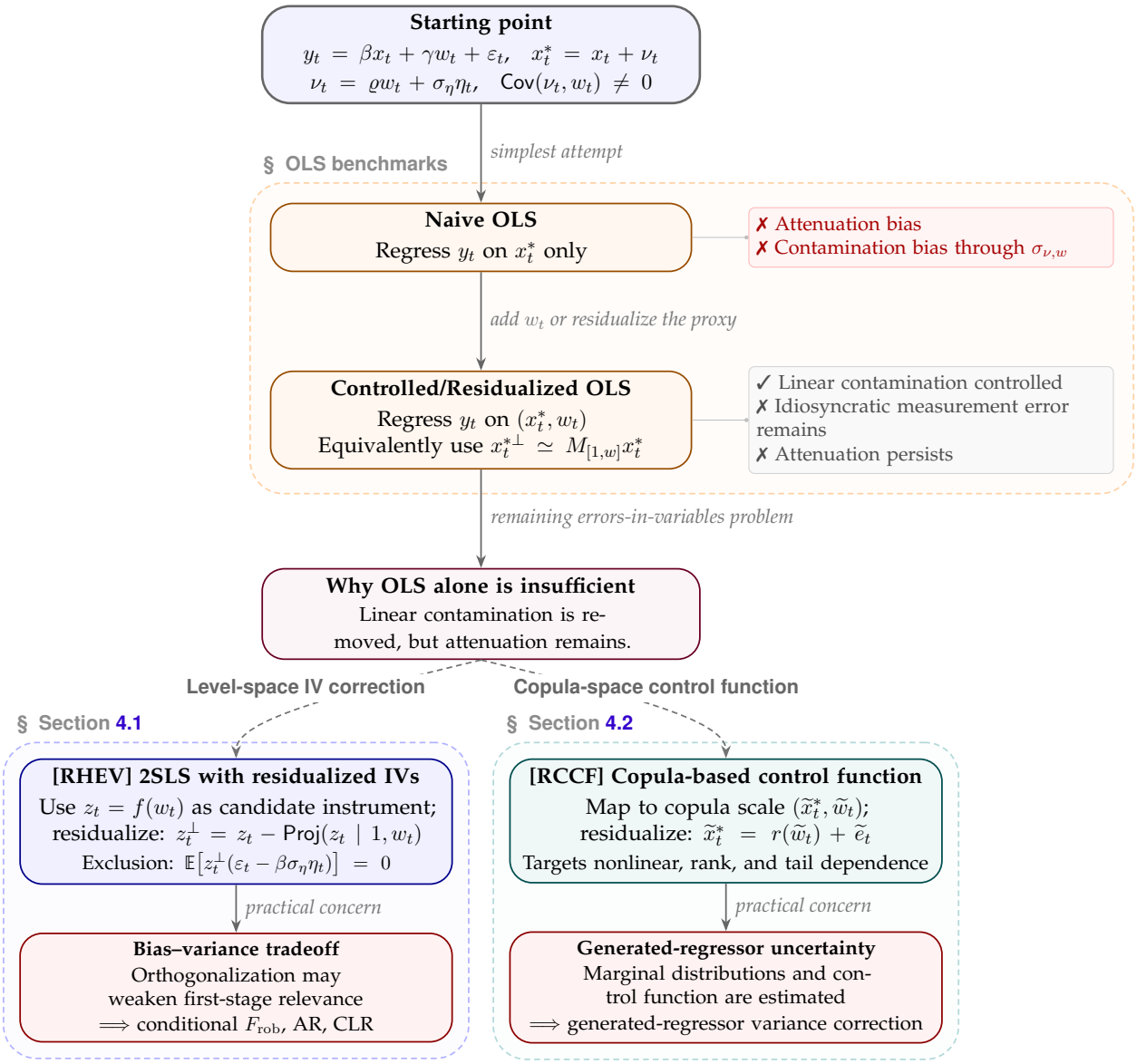


Figure 2: Roadmap of the Regression Problem and Correction Methods. *Notes:* The figure summarizes the logic of the regression problem and the two proposed correction strategies. Naive OLS suffers from both attenuation and contamination-induced omitted-variable bias. Controlled, or equivalently residualized, OLS removes the component of the proxy that is linearly predictable from the observable state variable, but the residualized proxy still contains idiosyncratic measurement error. RHEV approach addresses this remaining attenuation problem by using residualized high-orders of the observed exogenous variables (RHEVs) as instruments. The residualized copula-based control function (RCCF) approach provides a complementary dependence-based correction in copula space.

channels. This residual dependence is what our methods exploit: it is orthogonal to the linear contamination channel ϱw_t, yet still carries information about x_t. Method 1 (RHEV) accesses it through selected cross-moments, using a nonlinear transformation $z_t = f(w_t)$ that isolates particular higher-order cross-moments surviving the normalization; Method 2 (RCCF) accesses it through the joint distribution, using a copula for (x_t, w_t) that captures the dependence structure as a whole.

4.1 Method 1: 2SLS with Residualized High-Order Exogenous Variables

The first proposed method (RHEV) addresses the remaining attenuation problem by using transformations of the observable economic states as excluded instruments. The key step is to remove the component from the candidate instruments that is mechanically aligned with the linear contamination channel. This step is essential since the same state variable w_t that contaminates the proxy is also included as a control in the outcome equation.

RHEV Instrument Let $z_t = f(w_t)$ denote a candidate instrument, such as a centered polynomial transformation of w_t. For expositional simplicity, suppose $\dim(z_t) = 1$. Such a transformation can predict x_t through the higher-order cross-moments left unrestricted by the normalization $\mathsf{Cov}(x_t, w_t) = 0$, while the contamination channel ϱw_t remains linear in w_t. This is what makes z_t a candidate instrument: it can be relevant for the latent signal yet, once its linear component in w_t is removed, orthogonal to the contamination channel.

We residualize this candidate instrument with respect to the same control space used in the outcome equation,

$$z_t^{\perp} \equiv z_t - \mathsf{Proj}(z_t \mid 1, w_t),$$

which is reffered as a RHEV instrument variable. The orthogonalization (projection) is performed on the instrument rather than on the proxy. Its purpose is not to remove measurement error from the proxy, but to eliminate the component of the candidate instrument that is mechanically aligned with the included control w_t. The remaining variation can then serve as an excluded instrument for the residualized proxy.

The validity of this strategy requires an exclusion restriction. Since $x_t^{*\perp} = x_t + \sigma_\eta \eta_t$, the structural equation can be written as

$$y_t = \beta x_t^{*\perp} + \gamma w_t + \varepsilon_t - \beta \sigma_\eta \eta_t. \tag{15}$$

Thus, the relevant exclusion restriction is not only that $z_t^{\perp}$ be orthogonal to the proxy error, but that it be orthogonal to the full proxy-regression error, $\varepsilon_t - \beta \sigma_\eta \eta_t$. We impose this

condition as follows.

Assumption 5 (RHEV-IV Exclusion) *The linear contamination model, $x_t^* = x_t + \varrho w_t + \sigma_\eta \eta_t$, holds. Moreover, the idiosyncratic measurement component and the structural outcome error satisfy*

$$\mathbb{E}[\eta_t \mid w_t] = 0 \quad \textit{and} \quad \mathbb{E}[\varepsilon_t \mid w_t] = 0.$$

Finally, the direct effect of other economic states on the outcome is fully captured by the included linear control w_t in equation (12). In particular, excluded nonlinear transformations of w_t must not have a direct effect on y_t after the included linear control w_t is held fixed.

Assumption 5 strengthens the benchmark covariance restrictions used in the OLS analysis to conditional mean restrictions. The condition $\mathbb{E}[\eta_t \mid w_t] = 0$ ensures that functions of the observable state are orthogonal to the remaining text-measurement error after residualization. Likewise, $\mathbb{E}[\varepsilon_t \mid w_t] = 0$ ensures that nonlinear transformations of the observable state variable have no direct omitted effect on the outcome beyond the included linear control. Together these conditions deliver the exclusion restriction for RHEV estimator.

Lemma 3 (RHEV-IV Moment) *Suppose Assumption 5 holds. Then,*

$$\mathbb{E}[z_t^\perp \nu_t] = 0 \qquad \textit{and} \qquad \mathbb{E}\left[z_t^\perp(\varepsilon_t - \beta\sigma_\eta \eta_t)\right] = 0.$$

To see the first condition, note that

$$\mathbb{E}[z_t^\perp \nu_t] = \varrho \mathbb{E}[z_t^\perp w_t] + \sigma_\eta \mathbb{E}[z_t^\perp \eta_t] = 0.$$

The first term is zero by construction since $z_t^\perp$ is residualized with respect to $(1, w_t)$. The second term is zero since $z_t^\perp$ is a function of w_t and $\mathbb{E}[\eta_t \mid w_t] = 0$. Similarly, $\mathbb{E}[z_t^\perp \varepsilon_t] = 0$ follows because $z_t^\perp$ is a function of w_t and $\mathbb{E}[\varepsilon_t \mid w_t] = 0$. Hence, $z_t^\perp$ is orthogonal to the full proxy-regression error.

The second requirement is relevance. Projection removes the linear component of z_t aligned with w_t, but it must leave enough variation to predict the latent signal.

Assumption 6 (RHEV-IV Relevance) *The RHEV instrument $z_t^\perp$ satisfies $\mathbb{E}[z_t^\perp x_t] \neq 0$. The relevant second moments are finite, with $0 < \mathbb{E}[(z_t^\perp)^2] < \infty$ and $\mathbb{E}[x_t^2] < \infty$.*

Identification relies on residual nonlinear dependence between the latent and observable economic state variables. Since $z_t^\perp$ is constructed entirely from w_t,

$$\mathbb{E}[z_t^\perp x_t] = \mathbb{E}\left[z_t^\perp \mathbb{E}[x_t \mid w_t]\right].$$

Orthogonalization removes only the linear component associated with the contamination channel. The remaining nonlinear component must continue to predict the latent signal.

Residualized High-Order Exogeneous Variable (RHEV) Estimator Using the residualized high-order of observable state variables $z_t^{\perp}$ as an instrument, we define the RHEV estimator. For notational simplicity, all variables are centered and the constant is included in both projection spaces.

Definition 2 (RHEV Estimator) *For notational simplicity, variables are centered. Let $M_{[1,\boldsymbol{w}]}$ denote the residual-maker matrix from the sample projection on a constant and $\boldsymbol{w}$. The constant is included in the residualization step but suppressed from the displayed 2SLS matrices. We define*

$$\widehat{\boldsymbol{x}}^{*\perp} \equiv M_{[1,\boldsymbol{w}]}\boldsymbol{x}^*, \qquad \widehat{\boldsymbol{z}}^{\perp} \equiv M_{[1,\boldsymbol{w}]}\boldsymbol{z}, \text{ and } \widehat{D} \equiv \begin{pmatrix} \widehat{\boldsymbol{x}}^{*\perp} & \boldsymbol{w} \end{pmatrix}, \qquad \widehat{H} \equiv \begin{pmatrix} \widehat{\boldsymbol{z}}^{\perp} & \boldsymbol{w} \end{pmatrix},$$

where $\boldsymbol{x}^*, \boldsymbol{z}, \boldsymbol{w}$ denote the corresponding sample vectors. The hats on $\widehat{\boldsymbol{x}}^{*\perp}$ and $\widehat{\boldsymbol{z}}^{\perp}$ indicate sample residualization, not estimation of the latent signal $\boldsymbol{x}$.

The RHEV estimator is

$$\widehat{\boldsymbol{\theta}}^{\mathrm{RHEV}} = \left(\widehat{D}' P_{\widehat{H}} \widehat{D}\right)^{-1} \widehat{D}' P_{\widehat{H}} \boldsymbol{y}, \qquad P_{\widehat{H}} \equiv \widehat{H}\left(\widehat{H}'\widehat{H}\right)^{-1}\widehat{H}'.$$

We write $\widehat{\boldsymbol{\theta}}^{\mathrm{RHEV}} \equiv \left[\widehat{\beta}^{\mathrm{RHEV}}, \widehat{\gamma}^{\mathrm{RHEV}}\right]'$. Thus, $\widehat{\beta}^{\mathrm{RHEV}}$ is the RHEV estimator of the latent effect.

The RHEV-IV method jointly identifies the structural coefficient vector $\boldsymbol{\theta} \equiv [\beta, \gamma]'$. Unlike controlled OLS, which continues to use the noisy residualized proxy, RHEV keeps w_t as its own included instrument and uses $z_t^{\perp}$ only for the residualized proxy. Under Assumptions 5 and 6, the instrument (RHEV) is orthogonal to the proxy-regression error and relevant for the latent signal, yielding joint identification of β and γ.

Theorem 2 (Consistency and Asymptotic Normality of RHEV) *Under Assumptions 4, 5, and 6,*

$$\widehat{\boldsymbol{\theta}}^{\mathrm{RHEV}} \xrightarrow{p} \boldsymbol{\theta} \equiv [\beta, \gamma]', \text{ and } \sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{\mathrm{RHEV}} - \boldsymbol{\theta}\right) \rightsquigarrow \mathcal{N}(0, V^{\mathrm{RHEV}}),$$

where V^{RHEV} is the long-run variance.

The observed proxy contains three components: the latent signal x_t, the systematic contamination component ϱw_t, and the idiosyncratic text-measurement noise $\sigma_\eta \eta_t$. Controlled OLS and residualized OLS are numerically equivalent by the Frisch–Waugh–Lovell theorem. Both remove the component of the proxy that is linearly predictable from the observable economic state variable, leaving $x_t^{*\perp} = x_t + \sigma_\eta \eta_t$. Thus, controlled OLS eliminates

the linear contamination channel but remains attenuated since the residualized proxy is still measured with error.

The RHEV method adds a second correction step. It residualizes the candidate instrument against the same control space, $\widehat{\boldsymbol{z}}^{\perp} = M_{[1,\boldsymbol{w}]}\boldsymbol{z}$, and uses only the component of the residualized proxy $M_{[1,\boldsymbol{w}]}\boldsymbol{x}^*$ that is predicted by $\widehat{\boldsymbol{z}}^{\perp}$. Under the relevance condition $\mathbb{E}(z_t^{\perp} x_t) \neq 0$ and the exclusion condition that $z_t^{\perp}$ is orthogonal to the proxy-regression error, this IV-predicted variation identifies the latent effect β without attenuation from $\sigma_\eta \eta_t$.

Relevance Tradeoff Residualization strengthens the exclusion restriction by removing the component of the candidate instrument that is aligned with the contamination control w_t. However, the same residualization can weaken first-stage relevance. Therefore, Method 1 should be accompanied by conditional first-stage diagnostics and weak-IV robust inference, such as Anderson–Rubin or conditional likelihood-ratio procedures.

The scalar homoskedastic i.i.d. benchmark makes this tradeoff transparent. Suppose $z_t^{\perp} \in \mathbb{R}$. Then, the asymptotic variance can be represented by

$$V^{\mathrm{RHEV}} = \left(\sigma_\varepsilon^2 + \beta^2 \sigma_\eta^2\right) \begin{pmatrix} \sigma_x^{-2} \rho_{z^{\perp},x}^{-2} & 0 \\ 0 & \sigma_w^{-2} \end{pmatrix}, \tag{16}$$

where $\sigma_{z^{\perp}}^2 \equiv \mathbb{E}[(z_t^{\perp})^2]$, $\rho_{z^{\perp},x} = \frac{\sigma_{z^{\perp},x}}{\sigma_{z^{\perp}} \sigma_x}$, and $\sigma_{z^{\perp},x} \equiv \mathbb{E}[z_t^{\perp} x_t]$.[7] The remaining idiosyncratic text-measurement error $\sigma_\eta \eta_t$ is no longer a source of asymptotic bias after implementing RHEV as IV. However, it continues to inflate residual variance through $\sigma_\varepsilon^2 + \beta^2 \sigma_\eta^2$.

Statistical precision of our main interest (β) mainly depends on the predictive strength of the RHEV instrument, represented by $\sigma_x^{-2} \rho_{z^{\perp},x}^{-2}$. When the remaining nonlinear dependence between the surrounding economic environment and the latent signal is weak (small $\rho_{z^{\perp},x}$), estimation remains consistent but becomes imprecise. Consequently, empirical applications should routinely report conditional first-stage diagnostics together with weak-IV robust confidence intervals.

A Control-Function View Representing RHEV-IV approach as a control-function framework clarifies the connection between two complementary meothods. Under the linear contamination and the benchmark orthogonality condition $\mathsf{Cov}(x_t, w_t) = 0$, the population residualized proxy satisfies $x_t^{*\perp} = x_t + \sigma_\eta \eta_t$. The structural equation can therefore be written as

$$y_t = \beta x_t^{*\perp} + \gamma w_t + u_t^{\perp}, \qquad u_t^{\perp} \equiv \varepsilon_t - \beta \sigma_\eta \eta_t.$$

[7] Since $z_t^{\perp}$ is orthogonal to the idiosyncratic measurement error, $\mathbb{E}[z_t^{\perp} x_t] = \mathbb{E}[z_t^{\perp} x_t^{*\perp}]$.

Here, $u_t^{\perp}$ is the proxy-regression error that remains after the linear contamination component has been partialled out.

The 2SLS-IV first stage is

$$x_t^{*\perp} = \underbrace{\Pi z_t^{\perp}}_{\text{part of } x_t \text{ explained by } f(w_t)} + \big(x_t^{*\perp} - \Pi z_t^{\perp}\big), \qquad \mathbb{E}\left[z_t^{\perp}\big(x_t^{*\perp} - \Pi z_t^{\perp}\big)\right] = 0,$$

where $\Pi \equiv \frac{\mathbb{E}(z_t^{\perp} x_t^{*\perp})}{\mathbb{E}[(z_t^{\perp})^2]}$. The term $x_t^{*\perp} - \Pi z_t^{\perp}$ is the level-space first-stage residual: it is the component of the residualized proxy not explained by the RHEV instrument.

A control-function representation shows the dependence between $u_t^{\perp}$ and $x_t^{*\perp} - \Pi z_t^{\perp}$:

$$u_t^{\perp} = \phi^{\text{lev}}\big(x_t^{*\perp} - \Pi z_t^{\perp}\big) + \xi_t, \qquad \mathbb{E}\left[\big(x_t^{*\perp} - \Pi z_t^{\perp}\big)\xi_t\right] = 0.$$

The coefficient ϕ^{lev} is the level-space control-function coefficient. The corresponding residual-inclusion form is

$$y_t = \beta x_t^{*\perp} + \gamma w_t + \phi^{\text{lev}}\big(x_t^{*\perp} - \Pi z_t^{\perp}\big) + \xi_t.$$

The exclusion and relevance requirements have natural control-function interpretations. The IV exclusion condition is $\mathbb{E}(z_t^{\perp} u_t^{\perp}) = 0$. Since the first-stage residual is orthogonal to $z_t^{\perp}$, the above implies $\mathbb{E}(z_t^{\perp}\xi_t) = 0$. Together with $\mathbb{E}\left[\big(x_t^{*\perp} - \Pi z_t^{\perp}\big)\xi_t\right] = 0$, this gives

$$\mathbb{E}(x_t^{*\perp}\xi_t) = \Pi\mathbb{E}(z_t^{\perp}\xi_t) + \mathbb{E}\left[\big(x_t^{*\perp} - \Pi z_t^{\perp}\big)\xi_t\right] = 0.$$

Hence, the residual-inclusion equation identifies β in the same moment sense as the RHEV estimator.

The relevance condition is the nondegeneracy of the first stage:

$$\Pi \neq 0 \quad \Longleftrightarrow \quad \mathbb{E}(z_t^{\perp} x_t^{*\perp}) = \mathbb{E}(z_t^{\perp} x_t) \neq 0,$$

where the equality follows from $\mathbb{E}(z_t^{\perp}\eta_t) = 0$. If $\Pi = 0$, then the first-stage residual equals $x_t^{*\perp}$, so the residual-inclusion equation cannot separately identify the latent-effect coefficient β from the level-space control-function coefficient ϕ^{lev}.

This control-function view clarifies the bridge to the alternative method (RCCF). RHEV mothod uses a linear control function constructed from a level-space first stage:

$$\mathbb{E}\left[u_t^{\perp} \mid x_t^{*\perp}, w_t, x_t^{*\perp} - \Pi z_t^{\perp}\right] = \phi^{\text{lev}}\big(x_t^{*\perp} - \Pi z_t^{\perp}\big).$$

A more general representation would allow

$$\mathbb{E}\left[u_t^{\perp} \mid x_t^{*\perp}, w_t, x_t^{*\perp} - \Pi z_t^{\perp}\right] = h\left(x_t^{*\perp} - \Pi z_t^{\perp}\right),$$

where $h(\cdot)$ captures nonlinear conditional mean dependence between the residualized-proxy regression error and the level-space first-stage residual. This raises a further question: should the control function itself be constructed in level space? The next subsection answers this question by constructing the control function in copula space.

4.2 Method 2: Residualized Copula-Based Control Function Approach

The second method (RCCF) addresses the same remaining proxy-endogeneity problem from a different perspective. The previous RHEV method uses residualized nonlinear transformations of the observable state as excluded instruments. The alternative RCCF method instead constructs a generated control function from the dependence between the residualized proxy and the observable state on a copula scale. This approach is useful when the remaining dependence is nonlinear, rank-based, or concentrated in joint tail events, so that linear residualization in level space may fail to capture the relevant dependence.

As in constructing RHEV-IV, we first remove the linear contamination component. Under the linear contamination model and the benchmark orthogonality condition,

$$y_t = \beta x_t^{*\perp} + \gamma w_t + u_t^{\perp}, \quad , x_t^{*\perp} = x_t + \sigma_\eta \eta_t, \quad u_t^{\perp} \equiv \varepsilon_t - \beta \sigma_\eta \eta_t.$$

Thus, after controlling for w_t, the remaining proxy endogeneity comes from the idiosyncratic text-measurement component $\sigma_\eta \eta_t$. The RCCF method uses a copula-scale residual as a generated control for this remaining endogenous component. The method does not replace the need for an identifying restriction; rather, it replaces the excluded-instrument restriction of Method 1 with a copula control-function restriction.

Copula Control Function Let $F_{x^{*\perp}}$ and F_w denote the marginal cumulative distribution functions of the residualized proxy and the observable state. Construct the pseudo-observations

$$U_{x^{\perp},t} = F_{x^{*\perp}}(x_t^{*\perp}), \qquad U_{w,t} = F_w(w_t).$$

These probability-integral transformations remove the marginal distributions while preserving the rank dependence between $x_t^{*\perp}$ and w_t. We trim observations away from the unit-square boundaries so that the Gaussian quantile transformation is well-behaved.

Mapping the pseudo-observations to the Gaussian copula scale gives

$$\widetilde{x}_t^{*\perp} = \Phi^{-1}(U_{x^\perp,t}), \qquad \widetilde{w}_t = \Phi^{-1}(U_{w,t}),$$

where Φ^{-1} is the standard normal quantile function. The copula first stage models the conditional mean of the transformed residualized proxy given the transformed observable state:

$$\widetilde{x}_t^{*\perp} = r(\widetilde{w}_t) + \widetilde{e}_t, \qquad \mathbb{E}[\widetilde{e}_t \mid \widetilde{w}_t] = 0. \tag{17}$$

The residual $\widetilde{e}_t$ is not a decontaminated proxy in level space. Nor is it a transformed version of $x_t^{*\perp}$. Rather, it is the copula-scale innovation in the residualized proxy after removing the component predictable from the observable state on the copula scale.

To state the identifying restriction, recall that the residualized-proxy regression error is

$$u_t^\perp = \varepsilon_t - \beta\sigma_\eta\eta_t.$$

Method 2 assumes that the conditional mean dependence between this error and the residualized proxy is summarized by the copula residual $\widetilde{e}_t$. Specifically, we impose

$$u_t^\perp = \phi^{\text{cop}}\widetilde{e}_t + \zeta_t, \qquad \mathbb{E}[\zeta_t \mid w_t, \widetilde{e}_t] = 0. \tag{18}$$

Equivalently,

$$\mathbb{E}[u_t^\perp \mid w_t, \widetilde{e}_t] = \phi^{\text{cop}}\widetilde{e}_t.$$

This is the identifying restriction of RCCF method. It is not implied by the copula transformation alone. The copula transformation constructs a generated control; identification requires that this control be sufficient to span the conditional mean dependence between the residualized-proxy error and the residualized proxy.

The restriction also requires a nondegeneracy condition. The copula residual must be informative about the endogenous component of the residualized proxy, but it must not absorb all identifying variation in $x_t^{*\perp}$. Formally, define

$$S_{x^\perp} \equiv \mathbb{E}\left[\left(x_t^{*\perp} - \mathsf{Proj}(x_t^{*\perp} \mid 1, w_t, \widetilde{e}_t)\right)^2\right].$$

Identification of β requires $S_{x^\perp} > 0$. This condition is the control-function analog of a relevance condition: after conditioning on the observable state and the copula residual, the residualized proxy must retain independent variation with which to estimate the latent-effect coefficient.

Under (18), the augmented outcome equation is

$$y_t = \beta x_t^{*\perp} + \gamma w_t + \phi^{\text{cop}} \widetilde{e}_t + \zeta_t. \tag{19}$$

Conditional on the maintained control-function restriction and $S_{x^\perp} > 0$, the coefficient on $x_t^{*\perp}$ identifies the latent effect β. The coefficient γ remains the structural direct effect of the observable state because the linear contamination component has already been partialled out.

Assumption 7 (Copula Control-Function Validity and Regularity) *The marginal CDFs $F_{x^{*\perp}}$ and F_w are consistently estimated after trimming, and the Gaussian quantile transformation Φ^{-1} is continuously differentiable on the trimmed support. The copula-scale first-stage estimator $\widehat{r}$ is sufficiently regular so that the generated control-function moment admits a first-order stochastic expansion. The control-function restriction*

$$u_t^\perp = \phi^{\text{cop}} \widetilde{e}_t + \zeta_t, \qquad \mathbb{E}[\zeta_t \mid w_t, \widetilde{e}_t] = 0$$

holds. Finally,

$$S_{x^\perp} = \mathbb{E}\left[\left(x_t^{*\perp} - \mathsf{Proj}(x_t^{*\perp} \mid 1, w_t, \widetilde{e}_t)\right)^2\right] > 0,$$

and the second-step regressor matrix has a finite and nonsingular population second moment.

Assumption 7 combines generated-regressor regularity with the identifying control-function restriction. The regularity conditions ensure that estimation of the marginal distributions, the Gaussian copula-scale transformation, and the first-stage regression admits an asymptotic linear representation. Thus, the sampling variation from these first-step objects can be incorporated into the two-step variance. The identifying condition requires the copula residual $\widetilde{e}_t$ to absorb the conditional mean dependence between the residualized-proxy error and the residualized proxy, after conditioning on w_t. The nondegeneracy condition $S_{x^\perp} > 0$ ensures that the generated control does not absorb all variation in $x_t^{*\perp}$.

RCCF Estimator In sample, let $\widehat{x}_t^{*\perp}$ denote the residual from projecting x_t^* on a constant and w_t. Let $\widehat{F}_{x^{*\perp}}$ and $\widehat{F}_w$ be the empirical marginal CDFs of $\widehat{x}_t^{*\perp}$ and w_t. After trimming, construct

$$\widehat{U}_{x^\perp,t} = \widehat{F}_{x^{*\perp}}(\widehat{x}_t^{*\perp}), \quad \widehat{U}_{w,t} = \widehat{F}_w(w_t), \quad \widehat{\widetilde{x}}_t^{*\perp} = \Phi^{-1}(\widehat{U}_{x^\perp,t}), \quad \widehat{\widetilde{w}}_t = \Phi^{-1}(\widehat{U}_{w,t}).$$

The estimated copula residual is

$$\widehat{\widetilde{e}}_t \equiv \widehat{\widetilde{x}}_t^{*\perp} - \widehat{r}(\widehat{\widetilde{w}}_t).$$

We assume throughout that variables are centered or that an intercept is included in the second-step regression; to keep notation compact, the intercept is suppressed.

Definition 3 (RCCF Estimator) *The RCCF estimator as the OLS estimator from the augmented regression of y_t on $\widehat{R}_t$ is*

$$\widehat{\boldsymbol{\theta}}^{\mathrm{RCCF}} = \left(\sum_{t=1}^{n} \widehat{R}_t \widehat{R}_t'\right)^{-1} \left(\sum_{t=1}^{n} \widehat{R}_t y_t\right), \qquad \widehat{R}_t \equiv [\widehat{x}_t^{*\perp}, w_t, \widehat{\widetilde{e}}_t]'.$$

We write $\widehat{\boldsymbol{\theta}}^{\mathrm{RCCF}} \equiv [\widehat{\beta}^{\mathrm{RCCF}}, \widehat{\gamma}^{\mathrm{RCCF}}, \widehat{\phi}^{\mathrm{cop}}]'$. Thus, $\widehat{\beta}^{\mathrm{RCCF}}$ is the RCCF estimator of the latent effect, while $\widehat{\gamma}^{\mathrm{RCCF}}$ estimates the same structural direct effect of w_t as in Method 1.

Theorem 3 (RCCF Under the Maintained Control-Function Restriction) *Suppose Assumptions 4 and 7 hold. Then, under the maintained copula control-function restriction,*

$$\widehat{\boldsymbol{\theta}}^{\mathrm{RCCF}} \xrightarrow{p} \boldsymbol{\theta}^{\mathrm{cop}} \equiv [\beta, \gamma, \phi^{\mathrm{cop}}]', \text{ and } \sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{\mathrm{RCCF}} - \boldsymbol{\theta}^{\mathrm{cop}}\right) \rightsquigarrow \mathcal{N}(0, V^{\mathrm{RCCF}}).$$

The two-step variance V^{RCCF} accounts for sample residualization of the proxy, marginal-CDF estimation, the Gaussian copula-scale transformation, and the first-stage estimation of $r(\cdot)$.

Theorem 3 shows that RCCF estimator is consistent under the maintained copula control-function restriction. Identification does not come from excluded variation, as in Method 1. Instead, it comes from the assumption that the generated copula residual $\widetilde{e}_t$ is sufficient to capture the conditional mean dependence between the residualized-proxy disturbance $u_t^{\perp}$ and the residualized proxy after conditioning on the observable state. If this restriction is misspecified, the estimator generally converges to the corresponding augmented-regression pseudo-true value rather than to $(\beta, \gamma, \phi^{\mathrm{cop}})$.

Thus, both correction methods, RHEV and RCCF place the identifying burden in different places. RHEV requires credible residualized excluded variation. RCCF does not require excluded instruments, but it requires a credible copula control-function specification and the nondegeneracy condition $S_{x^{\perp}} > 0$.

Variance Tradeoff The scalar homoskedastic i.i.d. benchmark clarifies the variance tradeoff in Method 2. If the copula nuisance objects were known, the homoskedastic i.i.d.

variance of the coefficient on $x_t^{*\perp}$ in the augmented regression is

$$V_{\beta,\text{known}}^{\text{RCCF}} = \frac{\sigma_\zeta^2}{S_{x^\perp}}, \qquad S_{x^\perp} \equiv \mathbb{E}\left[\left(x_t^{*\perp} - \mathsf{Proj}(x_t^{*\perp} \mid 1, w_t, \widetilde{e}_t)\right)^2\right]. \tag{20}$$

Equivalently,

$$S_{x^\perp} = \mathsf{Var}(x_t^{*\perp})\left(1 - R^2_{x^{*\perp}|w,\widetilde{e}}\right),$$

where $R^2_{x^{*\perp}|w,\widetilde{e}}$ is the population R^2 from projecting $x_t^{*\perp}$ on a constant, w_t, and $\widetilde{e}_t$.[8] A small $S_{x^\perp}$ indicates that the generated control absorbs most of the residualized-proxy variation, making estimation of β imprecise.

When the copula nuisance components are estimated rather than known, the asymptotic variance is the variance of the combined second-step and first-step influence functions:

$$V_\beta^{\text{RCCF}} = \mathsf{Avar}\left(\psi_{\beta,t}^{\text{2nd}} + \psi_{\beta,t}^{\text{nuis}}\right).$$

The term $\psi_{\beta,t}^{\text{2nd}}$ is the influence function that would arise if $\widetilde{e}_t$ were observed, while $\psi_{\beta,t}^{\text{nuis}}$ collects the contribution from estimating the residualized proxy, the marginal distributions, the copula-scale transformation, and the first-stage regression $r(\cdot)$. The nuisance component typically increases sampling uncertainty, but its exact contribution depends on its covariance with the second-step score. Hence, inference should use a generated-regressor sandwich estimator or a bootstrap that re-estimates all first-step objects.

Comparison with RHEV-IV Approach Methods RHEV and RCCF solve the same residualized-proxy endogeneity problem using different identification strategies. RHEV is an instrumental-variables approach in level space. It identifies the latent effect using orthogonalized nonlinear transformations of observable states and therefore requires credible excluded variation after residualization.

RCCF is a control-function approach in copula space. Rather than relying on excluded instruments, it models dependence between the residualized proxy and the observable state through a generated copula residual. Consequently, it can accommodate nonlinear, rank-based, and tail dependence, although its validity depends on the correctness of the copula control-function restriction.

[8] When w_t and $\widetilde{e}_t$ are scalar, the same quantity can be written sequentially as $S_{x^\perp} = \mathsf{Var}(x_t^{*\perp})(1 - \rho^2_{x^{*\perp},w})(1 - \rho^2_{x^{*\perp},\widetilde{e}|w})$, where $\rho_{x^{*\perp},\widetilde{e}|w}$ is the partial correlation between $x_t^{*\perp}$ and $\widetilde{e}_t$ after both variables are residualized with respect to a constant and w_t. In the population residualized-proxy benchmark, $\rho_{x^{*\perp},w} = 0$, so this reduces to $S_{x^\perp} = \mathsf{Var}(x_t^{*\perp})(1-\rho^2_{x^{*\perp},\widetilde{e}|w})$. This expression should not be written with the unconditional correlation $\rho_{x^{*\perp},\widetilde{e}}$, since $\widetilde{e}_t$ is constructed to be orthogonal to $\widetilde{w}_t$ on the copula scale, not necessarily to w_t in level space.

Hence, the two methods involve different efficiency tradeoffs. The precision of RHEV estimator depends primarily on the strength of the residualized first stage. The precision of RCCF estimator depends on the residual variation in $x_t^{*\perp}$ after conditioning on both the observable state and the copula control function, as well as the sampling variability introduced by estimating the copula nuisance objects. Neither method uniformly dominates the other. The RHEV correction method is attractive when credible RHEV instruments are available, whereas RCCF correction is particularly appealing when writing-stage contamination is nonlinear, rank-based, or concentrated in joint tail events.

4.3 Inference, Diagnostics, and Reporting

The two correction methods have different inferential requirements because they place the identifying burden on different restrictions.

Inference for 2SLS-RHEV RHEV estimator is based on residualized-IV moment conditions and can therefore be implemented using IV inference with heteroskedasticity- and autocorrelation-consistent (HAC) covariance estimators. Since orthogonalization may weaken the residualized first stage, empirical applications should report first-stage diagnostics for the residualized regression of $x_t^{*\perp}$ on $z_t^{\perp}$, such as the partial-R^2 and the corresponding first-stage F-statistic. These statistics measure the strength of the identifying variation that remains after orthogonalization. Applications should also report weak-instrument robust procedures, such as Anderson–Rubin tests or conditional likelihood-ratio-type procedures when applicable.

Inference and Diagnostics for RCCF RCCF estimator relies on a generated copula control function and therefore requires inference that accounts for generated regressors. The asymptotic covariance matrix must incorporate sample residualization of the proxy, estimation of the marginal distributions, the Gaussian copula-scale transformation, and the first-stage estimation of $r(\cdot)$.

When the marginal distributions are estimated parametrically, a Murphy–Topel/Newey-type generated-regressor sandwich estimator can be used. In our baseline implementation, however, the marginal distributions are estimated by empirical CDFs. The rank transformation therefore introduces a nonparametric first-step component. For this reason, we recommend a block bootstrap that re-estimates all stages, including residualization, marginal-CDF estimation, construction of the copula-scale variables, estimation of $r(\cdot)$, and the augmented second-stage regression.

Diagnostics for RCCF should assess both the relevance of the generated control and the remaining identifying variation in the residualized proxy. In particular, empirical applications should report the strength and significance of the copula control-function coefficient $\widehat{\phi}^{\mathrm{cop}}$, as well as the nondegeneracy measure

$$\widehat{S}_{x^{\perp}} = \widehat{\mathsf{Var}}(\widehat{x}_t^{*\perp}) \left(1 - \widehat{R}^2_{x^{*\perp}|w,\widehat{\widetilde{e}}}\right).$$

A small value of $\widehat{S}_{x^{\perp}}$ indicates that the generated control absorbs most of the residualized-proxy variation, making estimation of β imprecise. Conversely, a weak or insignificant $\widehat{\phi}^{\mathrm{cop}}$ may indicate that the copula residual has limited ability to control the remaining proxy endogeneity, although it may also arise when the remaining proxy endogeneity is itself small and controlled OLS is already close to consistent.

These two diagnostics should be read jointly. A generated control that is strongly related to the remaining proxy endogeneity may also absorb substantial variation in $x_t^{*\perp}$, thereby reducing $\widehat{S}_{x^{\perp}}$ and increasing the variance of $\widehat{\beta}^{\mathrm{RCCF}}$. Conversely, a large $\widehat{S}_{x^{\perp}}$ by itself does not imply that RCCF is effective if $\widehat{\phi}^{\mathrm{cop}}$ is close to zero.

OLS Benchmarks as Diagnostics Finally, we recommend reporting naive OLS and controlled OLS in the linear-contamination benchmark, as diagnostic benchmarks rather than as correction methods. The comparison between naive OLS and controlled OLS illustrates the importance of linear writing-stage contamination. The comparison between controlled OLS and the proposed correction methods illustrates the importance of the remaining attenuation correction. Agreement between RHEV and RCCF methods is reassuring because the two estimators rely on different identifying restrictions. Disagreement is also informative, as it may signal weak RHEV instruments, misspecification of the copula control-function restriction, or sensitivity to the way nonlinear dependence is modeled.

5 Simulation Evidence

This section provides simulation evidence on the finite-sample behavior of the proposed procedures. The simulation is designed to capture three features of news-based uncertainty measures emphasized in the paper: persistent macroeconomic fundamentals, bad-news contamination in the observed text-based proxy, and nonlinear dependence between fundamentals and the proxy. Since latent uncertainty is unobserved and has arbitrary units in empirical applications, we distinguish between two empirical goals. The first is parameter recovery in a calibrated environment where the latent effect is known. The

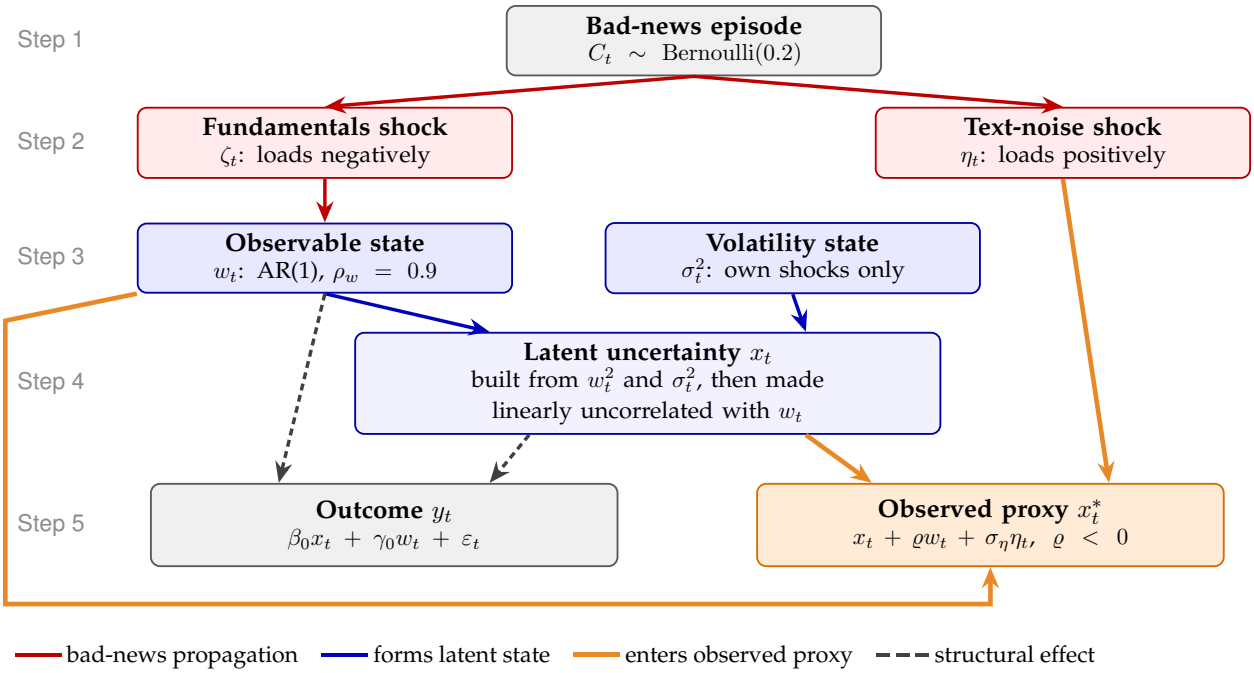


Figure 3: Causal Chain of the Data-Generating Process. *Notes:* Arrow colors mark the four channels; see Steps 1–5 in Section 5.1.

second is sign-oriented detection: whether an estimator recovers the expected negative association when the latent uncertainty effect is negative and avoids false negative findings in placebo designs. The tables below focus on conventional finite-sample performance measures, while the sign-oriented implications are discussed in the text.

5.1 Data-Generating Environment

The simulation design is built to reproduce three features of news-based uncertainty measures emphasized in the paper: a common bad-news episode that moves fundamentals and text measurement in opposite directions, a latent uncertainty state that is linearly uncorrelated with observable fundamentals yet nonlinearly dependent on them, and an observed proxy in which the same observable state enters both the measurement equation and the outcome equation. Figure 3 summarizes the resulting causal chain. We describe the environment in five steps, from primitive innovations to the outcome equation, and then separately describe the experimental designs we vary.

Step 1: Primitive innovations and the common bad-news episode. All randomness originates in four mutually independent innovation sequences,

$$e_{\zeta,t}, \quad e_{\eta,t}, \quad \xi_t, \quad \varepsilon_t,$$

which are independent over time and across components. In the symmetric heavy-tailed design, each is drawn from a standardized t_4 distribution with mean zero and unit variance.

In addition, a common episode indicator is drawn independently over time:

$$C_t \sim \text{Bernoulli}(q_C), \qquad q_C = 0.2,$$

This indicator is the only source of dependence between the fundamentals block and the text-measurement block, and it is what generates bad-news co-movement.

Step 2: The two opposite-signed bad-news channels. The episode indicator enters the fundamentals innovation and the text-measurement innovation with opposite signs:

$$\zeta_t = \mathsf{zscore}(e_{\zeta,t} - a_C(C_t - q_C)), \qquad \eta_t = \mathsf{zscore}(e_{\eta,t} + b_C(C_t - q_C)),$$

with $a_C = b_C = 1$. Standardization keeps both innovations mean-zero and unit-variance, so the episode changes their comovement without changing their scale. During a bad-news episode, ζ_t is shifted down and η_t is shifted up. Since ζ_t drives the observable state and η_t enters the text-measurement component of the proxy, adverse episodes simultaneously lower fundamentals and raise measured uncertainty. This is the simulation counterpart of the writing-stage mechanism in Section 2.2.

Step 3: The two state processes. The observable economic state follows a persistent first-order autoregression driven by the fundamentals innovation,

$$w_t = \rho_w w_{t-1} + \zeta_t, \qquad \rho_w = 0.9, \tag{21}$$

so that bad-news episodes have persistent effects on observable conditions. Separately, an autonomous stochastic-volatility state is driven only by its own innovation,

$$\ln \sigma_t^2 = (1 - \rho_h)\mu_{\ln \sigma^2} + \rho_h \ln \sigma_{t-1}^2 + \sigma_h \xi_t, \qquad \rho_h = 0.5, \quad \sigma_h = 0.9. \tag{22}$$

This second state is the part of uncertainty that is unrelated to the observable state, and it is what makes the relevance of the observable state for latent uncertainty less than complete.

Step 4: Latent uncertainty, linearly uncorrelated but nonlinearly dependent. Latent uncertainty mixes a fundamentals-volatility component with the autonomous volatility

component,

$$x_t^{\text{raw}} = w_w Z_{w^2,t} + (1 - w_w) Z_{\sigma^2,t},$$

where $Z_{w^2,t}$ and $Z_{\sigma^2,t}$ are the standardized versions of w_t^2 and σ_t^2. Building the first component from the *square* of the observable state is what creates dependence between latent uncertainty and observable conditions without creating a linear relationship. To impose the benchmark normalization exactly, we then residualize,

$$x_t = \mathsf{zscore}\big(M_{[1,w]} x_t^{\text{raw}}\big),$$

where $M_{[1,w]}$ is the sample residual-maker matrix from projecting on a constant and w_t. This guarantees $\mathsf{Cov}(x_t, w_t) = 0$ in every replication while leaving the nonlinear dependence intact. The design therefore realizes exactly the configuration that motivates our correction methods: linear orthogonality holds by construction, yet higher-order cross-moments between x_t and w_t remain informative.

Step 5: Contaminated proxy and outcome equation. The econometrician does not observe x_t. Instead, the observed news-based proxy adds a systematic contamination term and idiosyncratic text-measurement noise,

$$x_t^* = x_t + \varrho w_t + \sigma_\eta \eta_t, \qquad \varrho = -0.5, \quad \sigma_\eta = 0.9. \tag{23}$$

Since larger w_t denotes better fundamentals, $\varrho < 0$ is bad-news contamination: deteriorating fundamentals mechanically raise the observed uncertainty proxy. The outcome is generated by the structural equation

$$y_t = \beta_0 x_t + \gamma_0 w_t + \varepsilon_t, \qquad \gamma_0 = 0.45, \tag{24}$$

so that the observable state affects the outcome directly, in addition to entering the proxy through the contamination channel. This double role of w_t is precisely why controlling for it removes the contamination channel but cannot, by itself, repair the attenuation caused by $\sigma_\eta \eta_t$.

Experimental designs. Two features of the environment are varied across designs, holding everything above fixed. First, the relevance environment is governed by w_w, the weight on the fundamentals-volatility component in latent uncertainty. We report

$$w_w \in \{0.9,\ 0.5,\ 0.1\},$$

corresponding to a strong-relevance environment, an intermediate environment, and a weak-relevance environment in which latent uncertainty is mostly autonomous and therefore only weakly related to nonlinear transformations of the observable state.

Second, the outcome design is either an effect design or a placebo design:

$$\beta_0 = -0.3 \qquad \text{or} \qquad \beta_0 = 0.$$

The effect design evaluates whether a method detects and estimates a genuine negative latent uncertainty effect. The placebo design evaluates false sign detection, that is, how often a method reports a significantly negative association when latent uncertainty has no effect on the outcome at all.

We set $R = 1,000$ replications and $n = 300$ observations, discarding 400 burn-in observations, and report HAC standard errors with lag length $\lfloor n^{1/4} \rfloor$.

Estimators Compared We compare four main estimators. The first two are OLS benchmarks: OLS without fundamentals $(\widehat{\widetilde{b}})$ and OLS controlling for w_t $(\widehat{b})$. These benchmarks are included because empirical applications of news-based uncertainty measures often focus on whether the proxy exhibits the expected negative and statistically significant association with the outcome. The third estimator is Method 1 (RHEV), the estimator using quadratic and quartic transformations of fundamentals as excluded instruments. The fourth estimator is Method 2 (RCCF), the copula-based control-function estimator. All reported standard errors are HAC-based with lag length $\lfloor n^{1/4} \rfloor$. We set $R = 1000$, $n = 300$, and discard 400 burn-in observations.

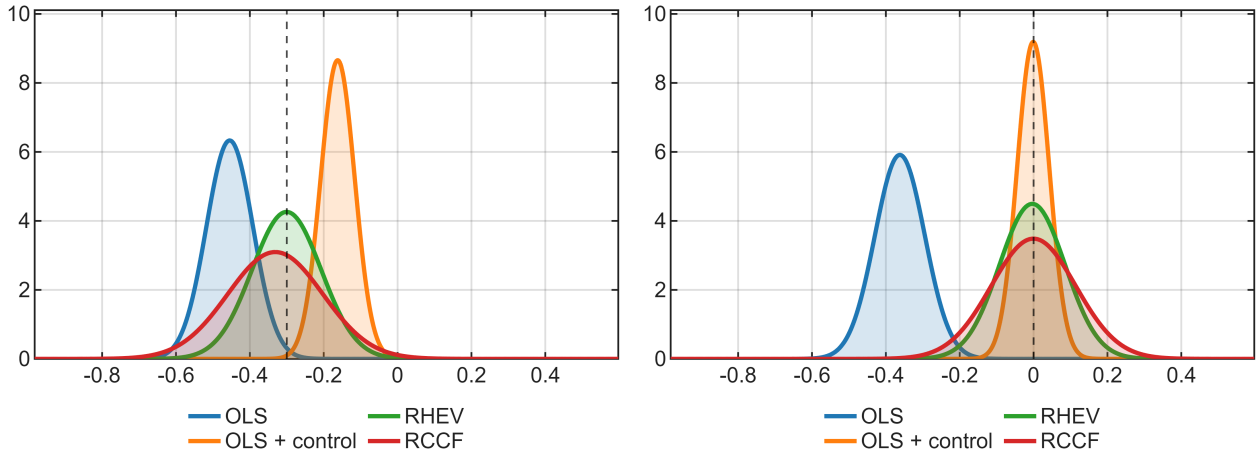
5.2 Simulation Results: Negative-Sign Detection Evidence

In applications using news-based uncertainty measures, researchers often place less emphasis on the numerical magnitude of the coefficient itself, since the latent uncertainty state is not directly observed and its scale is arbitrary. A central empirical question is instead whether the news-based proxy delivers a negative and statistically significant association with the outcome. Hence, we complement conventional finite-sample summaries with a sign-oriented diagnostic.

For each estimator, we report

$$\widehat{\Pr}\left(\widehat{\beta} < 0,\ 0 \notin \mathrm{CI}_{95}\right),$$

the Monte Carlo frequency with which the estimated coefficient is negative and its nominal



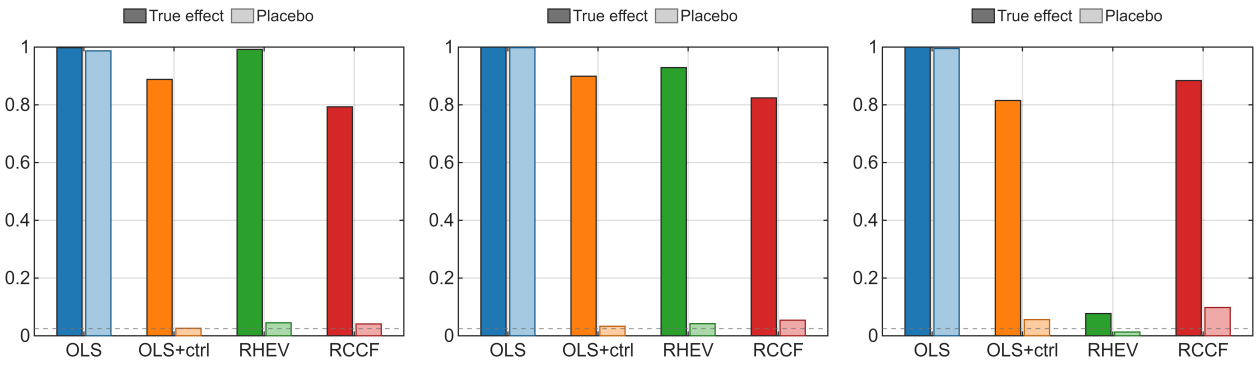
(a) True negative effect **(b)** No true effect

Figure 4: Simulated Distributions of Estimators. *Notes:* The figure plots sampling distributions of the four estimators in a representative intermediate relevance design ($w_w = 0.5$). The left panel reports the design with a true negative uncertainty effect, while the right panel reports the corresponding placebo design with no true effect.

95% confidence interval excludes zero. In the design with a true negative effect, $\beta_0 = -0.3$, this statistic measures correct negative detection power. In the placebo design ($\beta_0 = 0$), the same statistic measures false negative-sign discovery. Under correctly calibrated two-sided 95% inference with an approximately symmetric null distribution, the placebo benchmark is approximately 0.025.

Simulated Distributions Figure 4 illustrates the main mechanism behind the simulation. In the true-effect design, OLS without fundamentals is strongly shifted to the negative side. This may look like successful detection, but the placebo design shows that the same estimator also produces a strongly negative estimate when the latent uncertainty effect is exactly zero. Hence, the apparent power of OLS without fundamentals is not evidence that it recovers latent uncertainty; it is the empirical footprint of bad-news contamination.

Adding the fundamental control removes this spurious placebo effect and centers the estimator near zero in the placebo design. However, in the true-effect design, OLS with fundamentals is pulled toward zero because the proxy still contains idiosyncratic measurement error. RHEV is centered near the true negative effect in the nonzero design and near zero in the placebo design. RCCF also reduces the spurious negative bias, although it is more dispersed. Thus, the density figure summarizes the core distinction: controlling for fundamentals is necessary to avoid bad-news-driven false discoveries, but additional correction is needed to recover the latent uncertainty effect itself.



(a) Strong relevance **(b)** Intermediate relevance **(c)** Weak relevance

Figure 5: Negative-Effect Detection Across Relevance Environments. *Notes:* Each bar reports the Monte Carlo frequency with which an estimator produces a significantly negative uncertainty effect. Dark bars correspond to the design with a true negative effect, while light bars correspond to the placebo design with no true effect. The dashed horizontal line marks the nominal placebo benchmark of 0.025.

Sign-Oriented Diagnostics Figure 5 reports the same logic in terms of negative-effect detection. OLS without fundamentals yields a significantly negative effect with high probability across all relevance environments. However, the placebo bars are almost as large as the true-effect bars. Hence, OLS without fundamentals is an apparently powerful but misleading sign detector: it cannot distinguish a true latent uncertainty effect from the bad-news fundamentals channel. This is the main false-discovery problem created by contaminated news-based proxies.

The OLS-with-fundamentals benchmark has the opposite profile. Once fundamentals are included, placebo false discoveries fall close to the nominal negative-tail benchmark. This confirms that the observed bad-news channel is the source of the spurious negative coefficient in OLS without fundamentals. At the same time, OLS with fundamentals has lower detection power in the true-effect design because it still relies on a noisy proxy for latent uncertainty. Thus, conventional controls are valuable descriptive benchmarks, but they do not solve the remaining measurement-error problem.

The two proposed corrections address this remaining problem in different ways. When the relevance of fundamentals-volatility transformations is strong or intermediate, RHEV delivers high detection power in the true-effect design while keeping placebo false discoveries close to the nominal benchmark. This is the cleanest environment for RHEV: nonlinear transformations of fundamentals contain enough residualized first-stage information to help identify the latent uncertainty component after controlling for the linear bad-news channel. In the weak-relevance environment, however, RHEV becomes much less informative. Its placebo behavior remains conservative, but its true-effect detection power falls sharply, reflecting weak first-stage information rather than successful recovery.

RCCF provides a complementary correction. It retains substantial negative-effect detection even when the residualized high-order signal is weak, consistent with its use of nonlinear rank-based dependence between the residualized proxy and fundamentals. This additional detection power comes with a cost: the placebo bars are larger than those of RHEV, especially in the weak-relevance design. RCCF should therefore be interpreted as a useful complementary procedure rather than as a uniform replacement for RHEV.

6 Correction Method Choice: A Calibration Framework

Empirical practitioners face the challenge of choosing between two primary correction strategies: RHEV and RCCF estimators. Rather than viewing these approaches as competing, this section shows that they are complementary identification strategies tailored to different *relevance environments*. Because the underlying weight of fundamentals in latent uncertainty (w_w) and the proxy-noise scale (σ_η^2) are unobserved, we formalize a simulation-driven calibration framework that maps observable first-stage diagnostics into a practical method-choice decision.

6.1 Method Ranking and Relevance Environments

Section 5.2 shows that bad-news contamination creates two distinct problems. First, an estimator can falsely detect a negative uncertainty effect when the true effect is zero. Second, after controlling for the direct fundamentals channel, an estimator can still fail to detect a true negative effect since the proxy remains measured with error or because the identifying variation is weak. The placebo design is useful for diagnosing the first problem. In this subsection, we focus on the second problem: conditional on using estimators that control for or correct the fundamentals channel, which estimator is most informative for detecting a true negative uncertainty effect?

Simulation Setup In the simulation design, the relevance environment is governed by w_w, the weight on the fundamentals-volatility component in latent uncertainty. When w_w is small, latent uncertainty is mostly autonomous and high-order transformations of fundamentals provide little first-stage information. When w_w is large, these transformations become strongly informative for the residualized proxy. To trace how method performance changes with this environment, we run a dense-grid simulation over

$$w_w \in \{0.025, 0.050, \ldots, 0.975\}.$$

Table 1: Full-Factorial Calibration Design

Parameter	Interpretation	Values used in the full-factorial grid
w_w	Fundamentals relevance weight	$\{0.025, 0.050, \ldots, 0.975\}$
β_0	True latent uncertainty effect	$\{-0.45, -0.30, -0.15\}$
γ_0	Direct fundamentals effect	$\{0.25, 0.45, 0.65\}$
ρ_w	Fundamentals persistence	$\{0.70, 0.95\}$
ϱ	Bad-news loading in proxy	$\{-0.70, -0.30\}$
σ_η	Idiosyncratic proxy-noise scale	$\{0.60, 0.75, 0.90, 1.05, 1.20\}$

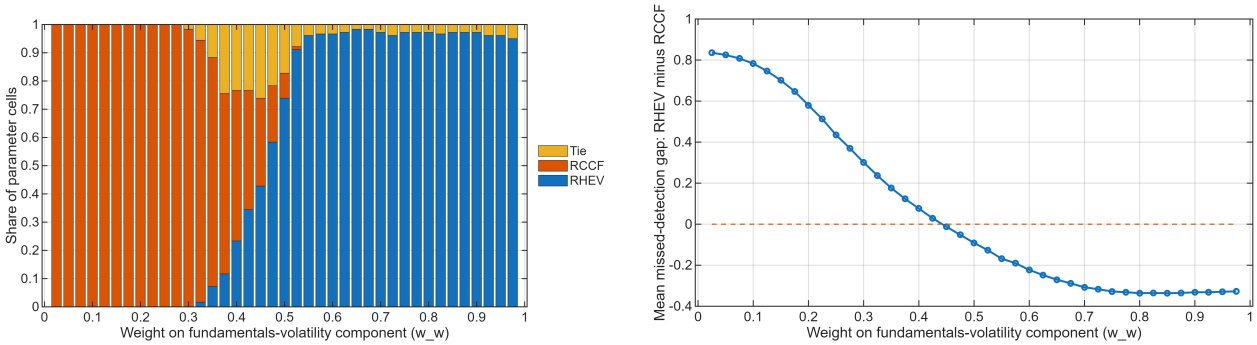
Notes: For each value of w_w, the simulation considers all combinations of the listed values of β_0, γ_0, ρ_w, ϱ, and σ_η. This gives $39 \times 3 \times 3 \times 2 \times 2 \times 5 = 7{,}020$ full-factorial cells. The parameter w_w indexes the relevance environment: larger values imply that residualized high-order transformations of fundamentals contain more information about latent uncertainty.

Rather than relying on a single baseline parameterization, we evaluate the methods over a reduced full-factorial grid in Table 1. Specifically, for each value of w_w, we consider all combinations of selected values of β_0, γ_0, ρ_w, ϱ, and σ_η drawn from the robustness ranges used in that table.

The comparison here focuses on the two correction methods, the RHEV estimator and the RCCF estimator. The purpose is not to revisit whether controlled OLS can sometimes perform well as a descriptive benchmark, but to determine which of the two correction-based approaches is preferred across relevance environments.[9] For each parameter cell and each correction method, we compute the missed-detection probability under the true negative-effect designs. We then summarize, for each w_w, the share of parameter cells in which RHEV has the lower missed-detection probability, the share in which RCCF has the lower missed-detection probability, and the share of ties. We also compute the average missed-detection gap, defined as the RHEV missed-detection probability minus the RCCF missed-detection probability. A positive value therefore favors RCCF, while a negative value favors RHEV.

Correction-Method Ranking and Relevance Environments Figure 6 shows that the preferred correction method changes systematically with the relevance environment. When w_w is low, RCCF dominates almost all parameter cells. In this region, the residualized first stage is weak, so RHEV has little identifying variation to exploit, whereas RCCF can still use nonlinear dependence between the proxy and fundamentals. This is reflected

[9] Naive OLS is excluded because its apparent detection power is contaminated by false discovery in the placebo design. OLS with fundamentals is also omitted from the present comparison since the question here is narrower: conditional on using a correction method, should one prefer RHEV or RCCF?



(a) Correction-only winner shares

(b) Average missed-detection gap

Figure 6: Correction-Method Ranking Across Relevance Environments. *Notes:* Panel (a) reports, for each value of w_w, the share of full-factorial parameter cells in which RCCF has a lower missed-detection probability than RHEV, the share in which RHEV has a lower missed-detection probability than RCCF, and the share of ties. Panel (b) reports the average missed-detection gap, defined as the missed-detection probability of RHEV minus that of RCCF, averaged across the same parameter cells. Positive values favor RCCF; negative values favor RHEV.

in Panel (a) by the large RCCF winner share and in Panel (b) by the strongly positive RHEV-minus-RCCF missed-detection gap.

As w_w increases, the RCCF advantage fades and the ranking passes through a transition region. In this middle range, both methods can be competitive: Panel (a) shows a mixture of RCCF wins, RHEV wins, and ties, while Panel (b) shows the average gap moving toward zero. Once w_w becomes sufficiently large, the ranking reverses. RHEV becomes the preferred correction in the overwhelming majority of parameter cells, and the average gap turns negative. This is the environment in which high-order transformations of fundamentals contain enough residualized first-stage information for the RHEV strategy to be most compelling.

Taken together, the figure supports a relevance-based correction rule: low relevance favors RCCF, high relevance favors RHEV, and the intermediate region is a transition zone in which both corrections should be reported. This is the ranking pattern that the calibration exercise seeks to recover from empirical first-stage diagnostics.

Recovering Relevance Environments from First-Stage Diagnostics The practical limitation of Figure 6 is that w_w is a simulation parameter and is not observed in empirical applications. Hence, we use an observable first-stage relevance score to locate the empirical application in a w_w-like relevance environment. Let

$$\kappa_{\mathrm{FS}} \equiv \widehat{R}^2_{\mathrm{FS}},$$

Table 2: Recovering the w_w-like relevance environment

Calibration specification	Proxy-noise adjustment	Fit R^2
Pooled linear κ_{FS}	No	0.875
Noise-adjusted linear κ_{FS}	Yes	0.969
Separate linear fits by σ_η	Group-specific	0.968–0.969

Notes: $\kappa_{\mathrm{FS}} \equiv \widehat{R}^2_{\mathrm{FS}}$ is the residualized first-stage partial R^2 from regressing $M_Q x_t^*$ on the residualized excluded instruments $M_Q Z_t$. The dependent variable is the simulation parameter w_w. The pooled and noise-adjusted specifications use all 7,020 full-factorial cells. The separate fits use five σ_η groups, with 1,404 cells in each group. The noise-adjusted specification allows the intercept and slope on κ_{FS} to vary with the proxy-noise ratio σ_η^2.

where $\widehat{R}^2_{\mathrm{FS}}$ is the residualized first-stage partial R^2 from regressing $M_Q x_t^*$ on the residualized excluded instruments $M_Q Z_t$. We begin with the diagnostic regression

$$w_w = a_0 + a_1 \kappa_{\mathrm{FS}} + e. \tag{25}$$

This equation is not a structural model and does not estimate w_w as an economic parameter. Rather, it provides a scale-free way to map an observable first-stage diagnostic into the weak, intermediate, and strong relevance environments studied in the simulation.

Table 2 shows that the observable first-stage diagnostic recovers the simulation-relevant environment well. The pooled linear mapping from κ_{FS} to w_w explains about 87.5% of the variation in the full-factorial design. This is already high, given that the design varies not only the relevance environment but also the outcome effect, the fundamentals channel, fundamentals persistence, bad-news contamination, and proxy-noise scale.

The remaining variation is mainly due to the proxy-noise environment. Once the mapping is allowed to vary with σ_η^2, the recovery fit rises to about 96.9%. Similarly, estimating the linear mapping separately by σ_η yields fits of about 96.8–96.9% in each proxy-noise group. This confirms that κ_{FS} is the relevant empirical diagnostic, while also showing why a single universal coefficient mapping is not appropriate.[10]

The robustness exercise delivers an important simplification. Across changes in the outcome-effect size β_0, the fundamentals effect γ_0, fundamentals persistence ρ_w, and the strength of linear bad-news contamination ϱ, the κ_{FS}-based ranking of relevance environments is highly stable. The parameter that materially shifts the calibration mapping

[10]We do not interpret the coefficients in the diagnostic regression as structural parameters. Their role is only to translate an observable residualized first-stage partial R^2 into a w_w-like relevance environment. The HAC first-stage statistic remains useful as a supplementary weak-relevance diagnostic, but it is not part of the main calibration rule.

is the idiosyncratic proxy-noise scale σ_η. This pattern is exactly what the residualized measurement equation implies:

$$M_Q x_t^* = M_Q x_t + \sigma_\eta M_Q \eta_t. \tag{26}$$

The outcome parameters affect detection power and false-discovery behavior, and the linear bad-news component is removed by projection. By contrast, σ_η^2 directly changes the fraction of residualized proxy variation that can be explained by the projected fundamentals instruments. Thus, empirical calibration should be based on the observed first-stage strength κ_{FS}, while recognizing that the same value of κ_{FS} can correspond to different w_w-like environments depending on the proxy-noise environment.

6.2 Proxy-Noise Sensitivity and Practical Implementation

In empirical applications, translating observable first-stage strength into a method-choice decision requires addressing an underlying identification challenge: the proxy-noise variance σ_η^2 is not directly observed from the final proxy series alone. In this subsection, we formalize how researchers can discipline proxy noise through sensitivity analysis, introduce an outcome-based diagnostic, and outline a step-by-step decision protocol.

Accounting for Unobserved Proxy Noise Variance The calibration results create a practical difficulty: the proxy-noise variance σ_η^2 is not directly observed and is generally not point-identified from the final proxy time series alone. The observed residualized proxy $M_Q x_t^*$ cannot uniquely reveal how much of its variation comes from latent uncertainty rather than idiosyncratic proxy noise. Thus, a low value of κ_{FS} may indicate either weak nonlinear relevance of fundamentals for latent uncertainty or a high-noise proxy. We therefore avoid a single universal conversion formula from κ_{FS} to w_w and tailor the empirical implementation to the available information about proxy construction and attenuation.

To account for the proxy-noise environment, we allow the calibration mapping to depend directly on the simulation-normalized proxy-noise scale σ_η^2. The full-factorial calibration suggests that a mapping based only on κ_{FS} is too restrictive, but that the relationship becomes highly stable once σ_η^2 is included. We therefore use the following simulation-calibrated conversion rule:

$$\widehat{w}_w^{\mathrm{cal}}(\sigma_\eta^2) = \left(0.1334 - 0.0070\sigma_\eta^2\right) + \left(0.7556 + 0.7177\sigma_\eta^2\right)\kappa_{\mathrm{FS}}. \tag{27}$$

This noise-adjusted mapping has a fitted R^2 of 0.9689 in the full-factorial calibration design.

The calibrated index is truncated to the unit interval:

$$\widehat{w}_w^{\mathrm{cal,tr}}(\sigma_\eta^2) = \min\left\{1, \max\left\{0, \widehat{w}_w^{\mathrm{cal}}(\sigma_\eta^2)\right\}\right\}. \tag{28}$$

The point of this expression is not that σ_η^2 is always observed. Rather, it makes explicit which feature of the proxy-construction environment shifts the mapping from observed first-stage strength to the underlying relevance environment.

Hence, the goal is not to estimate σ_η^2 as a structural parameter in every application. Rather, the goal is to discipline the proxy-noise environment as much as the available information allows. In the main text, we focus on the most common setting in which the researcher observes only the final proxy series x_t^*. Richer settings with construction metadata or raw text units provide additional ways to discipline proxy noise; we discuss those extensions in Appendix D.

When only the final proxy is available, the researcher can compute the observable first-stage relevance score κ_{FS}, but cannot separately identify the proxy-noise scale σ_η^2 from the final time series alone. A low value of κ_{FS} may therefore reflect weak nonlinear relevance of fundamentals for latent uncertainty, a high-noise proxy, or both. The empirical implementation should therefore report a sensitivity analysis over plausible values of σ_η^2. We report sensitivity over the proxy-noise scale used in the full-factorial calibration,

$$\sigma_\eta^2 \in \{0.36,\ 0.5625,\ 0.81,\ 1.1025,\ 1.44\}.$$

For each value of σ_η^2, the researcher computes

$$\widehat{w}_w^{\mathrm{cal,tr}}(\sigma_\eta^2)$$

using the noise-adjusted calibration rule above and checks whether the implied relevance classification changes. If the classification is stable across the grid, method choice is robust to proxy-noise uncertainty. If it is not stable, the application should report the corresponding range of plausible correction methods rather than selecting a single estimator mechanically.

Outcome-Based Proxy-Noise Diagnostic A further model-implied reliability diagnostic is available when a corrected estimate of β is credible. In practice, we use this diagnostic only when the RCCF estimates are stable across alternative copula specifications and pass

the placebo checks.[11] If the corrected estimates are imprecise or unstable across copulas, the attenuation ratio is not informative, and we rely instead on the proxy-noise sensitivity grid described above.

Under a classical residual measurement-error interpretation, OLS with fundamentals is attenuated toward zero relative to a consistent corrected estimate (i.e., $\operatorname{plim} \widehat{b} = \beta \frac{\sigma_x^2}{\sigma_x^2+\sigma_\eta^2}$). If the RCCF estimator is consistent for β, so that $\operatorname{plim} \widehat{\beta}^{\text{RCCF}} = \beta$, then

$$\operatorname{plim}\left(\frac{\widehat{\beta}^{\text{RCCF}}}{\widehat{b}} - 1\right) = \frac{\sigma_\eta^2}{\sigma_x^2}. \tag{29}$$

This ratio corresponds to the proxy-noise scale σ_η^2. This motivates the outcome-based proxy-noise diagnostic:

$$\widehat{\sigma}_{\eta,out}^2 = \frac{\widehat{\beta}^{\text{RCCF}}}{\widehat{b}} - 1. \tag{30}$$

We use this diagnostic only when the two estimates have the same sign and the OLS-with-fundamentals estimate is smaller in absolute value:

$$0 < \frac{\widehat{b}}{\widehat{\beta}^{\text{RCCF}}} \leq 1.$$

This condition is not an additional substantive restriction in the cases where the diagnostic is most useful. It is exactly the attenuation pattern expected when both estimators point in the same direction, but the OLS-with-fundamentals estimate is pulled toward zero by residual measurement error. Values outside this range indicate that the maintained attenuation model, the RCCF restriction, or finite-sample precision is not adequate for using the ratio mechanically. We therefore interpret $\widehat{\sigma}_{\eta,out}^2$ as a model-implied reliability diagnostic, not as an independent estimate of proxy noise.

Figure 7 validates the outcome-based calibration diagnostic in the full-factorial simulation. The calibrated relevance index tracks the true w_w closely, with most points lying near the 45-degree line. The diagnostic is somewhat noisier than the oracle calibration that uses the true σ_η^2, but it preserves most of the relevance classification information needed for method choice. This supports using $\widehat{\sigma}_{\eta,out}^2$ as a model-implied reliability diagnostic when the admissible attenuation pattern holds.

[11] When calibrated relevance is sufficiently strong, the RHEV estimate provides an additional corrected benchmark. Because RHEV is sensitive to weak first-stage relevance, we use it as corroborating evidence rather than as a mechanical substitute for the RCCF-based ratio.

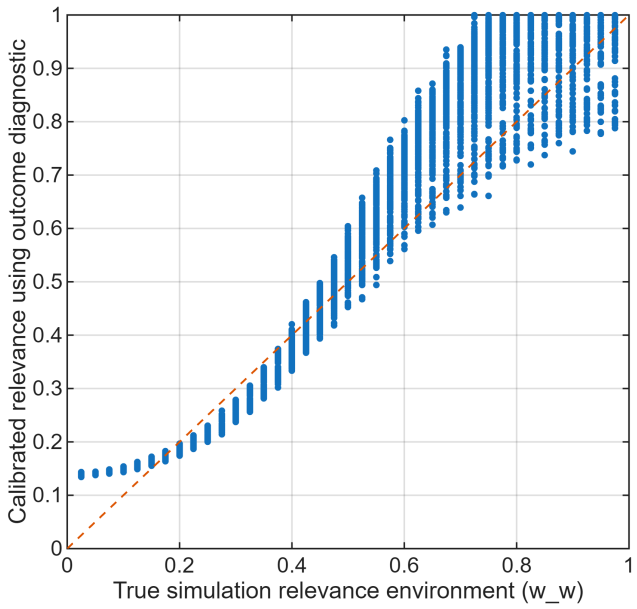


Figure 7: Outcome-Based Validation of the Calibrated Relevance Index. *Notes:* The horizontal axis is the true simulation relevance environment w_w, and the vertical axis is the calibrated relevance index obtained by inserting the outcome-implied proxy-noise diagnostic $\widehat{\sigma}^2_{\eta,out} = \widehat{\beta}^{\mathrm{RCCF}}/\widehat{b} - 1$ into the noise-adjusted calibration rule. The dashed line is the 45-degree line. The figure uses cell-level Monte Carlo averages from the 7,020 full-factorial calibration cells. All cells satisfy the admissible attenuation condition $0 < \widehat{b}/\widehat{\beta}^{\mathrm{RCCF}} \leq 1$. The calibrated relevance index has MAE 0.076, RMSE 0.098, and relevance-class agreement 0.861. As an oracle benchmark, using the true simulation value of σ^2_η gives MAE 0.042, RMSE 0.050, and relevance-class agreement 0.891.

Practical Decision Protocol for Empirical Applications Taken together, Figure 6 and the noise-adjusted calibration rule provide a relevance-based empirical implementation. In final-proxy-only applications, the researcher first estimates the residualized first stage and computes κ_{FS}. Since σ^2_η is not identified from the final proxy alone, the researcher then evaluates $\widehat{w}^{\mathrm{cal,tr}}_w(\sigma^2_\eta)$ over the proxy-noise sensitivity grid, using the outcome-based reliability diagnostic only when its maintained assumptions are credible. This produces a range of calibrated relevance environments rather than a single mechanical classification.

The resulting relevance range guides, but does not mechanically determine, method choice. Low calibrated relevance favors RCCF as the main corrected evidence, provided the copula-based estimates are stable and pass placebo checks. High calibrated relevance favors RHEV, because the residualized high-order instruments then contain enough first-stage information to be useful. Intermediate relevance calls for reporting both corrections. Thus, the calibration exercise organizes the empirical evidence around first-stage relevance, proxy-noise uncertainty, false-discovery diagnostics, and robustness across correction methods, rather than selecting an estimator mechanically.

7 Concluding Remarks

This paper develops a framework for understanding and correcting a fundamental measurement problem in empirical research with text-based proxy regressors. We show that text-derived proxies generally do not behave like classical noisy measurements. Because language is generated through context-dependent semantic-field activation and lexical co-selection, text-based measures systematically inherit information about observable economic states alongside the latent variables they are intended to represent. Consequently, empirical estimates based on these proxies may confound the effects of latent concepts with the effects of the economic environment embedded in the underlying text.

We formalize this contamination and develops econometric procedures that separate the latent component from the economic state-driven component of text-based measures. Rather than treating text-derived variables as given objects, we reframe their construction as a measurement problem requiring econometric correction. This perspective provides a more credible basis for causal inference and counterfactual analysis. For example, when news-based uncertainty measures are used as regressors, our methods recover a component that more closely corresponds to the theoretical notion of pure second-moment uncertainty, thereby enabling cleaner empirical tests of theories of investment, consumption, asset pricing, and other applications in which uncertainty plays a central role.

Our results can extend beyond textual analysis to NLP, ML, and AI methods that transform unstructured data into structured variables for empirical analysis. We highlight a fundamental distinction between two sources of measurement error in such variables. The first arises from imperfect statistical learning and representation, which advances in ML and AI continue to reduce. The second originates from the data-generation process itself: unstructured data reflect both the latent object of interest and the observable environments in which they are created. Improvements in representation learning do not eliminate this second source of error and may instead recover it more accurately. Consequently, valid inference with machine-generated variables requires accounting not only for algorithmic data-generating process error but also for systematic contamination inherited from the underlying data-generation process.

We therefore view this paper as an introductory framework for an econometric theory of representation learning for empirical economics. As empirical research increasingly relies on NLP, ML, and AI to transform unstructured data into quantitative measures, understanding the statistical properties of these learned representations—and separating latent signals from systematic contamination—will become an important prerequisite for credible empirical inference.

References

Alfaro, Ivan, Nicholas Bloom, and Xiaoji Lin, "The finance uncertainty multiplier," *Journal of Political Economy*, 2024, *132* (2).

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, July 2016, *131* (4), 1593–1636.

Battaglia, Laura, Timothy Christensen, Stephen Hansen, and Szymon Sacher, "Inference for Regression with Variables Generated by AI or Machine Learning," Cowles Foundation Discussion Papers 2421, Cowles Foundation for Research in Economics, Yale University 2025.

Baumeister, Roy F., Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs, "Bad is Stronger than Good," *Review of General Psychology*, 2001, *5* (4), 323–370.

Burtch, Gordon, Edward McFowland, Mochen Yang, and Gediminas Adomavicius, "EnsembleIV: Creating Instrumental Variables from Ensemble Learners for Robust Statistical Inference with ML- Generated Variables," *Management Science*, 2026, *Forthcoming*.

Caldara, Dario and Matteo Iacoviello, "Measuring Geopolitical Risk," *American Economic Review*, April 2022, *112* (4), 1194–1225.

_ , _ , Patrick Molligo, Andrea Prestipino, and Andrea Raffo, "The Economic Effects of Trade Policy Uncertainty," *Journal of Monetary Economics*, 2020, *109*, 38–59.

Calomiris, Charles W. and Harry Mamaysky, "How news and its context drive risk and returns around the world," *Journal of Financial Economics*, 2019, *133* (2), 299–336.

Calomiris, Charles W, Harry Mamaysky, and Ruoke Yang, "Measuring the Cost of Regulation: A Text-Based Approach," Working Paper 26856, National Bureau of Economic Research March 2020.

Damstra, Alyt and Mark Boukes, "The Economy, the News, and the Public: A Longitudinal Study of the Impact of Economic News on Economic Evaluations and Expectations," *Communication Research*, 2021, *48* (1), 26–50.

Fajgelbaum, Pablo D., Edouard Schaal, and Mathieu Taschereau-Dumouchel, "Uncertainty Traps," *The Quarterly Journal of Economics*, 2017, *132* (4), 1641–1692.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera, "The Voice of Monetary Policy," *American Economic Review*, February 2023, *113* (2), 548–84.

Hoberg, Gerard and Gordon Phillips, "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, 2016, *124* (5), 1423–1465.

Husted, Lucas, John Rogers, and Bo Sun, "Monetary policy uncertainty," *Journal of Monetary Economics*, 2020, *115*, 20–36.

Kahneman, Daniel and Amos Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 1979, *47* (2), 263–291.

Kim, Chan and Daisoon Kim, "Public Information on Overseas Buyers for Export Promotion," Working Paper, International Monetary Fund 2026.

Kim, Daisoon and Sunhyung Lee, "Multinational Investment Activity under Policy Uncertainty in Host and Competing Countries," *Journal of International Economics*, 2026, *161*, 104252.

Lyons, John, *Semantics*, Cambridge: Cambridge University Press, 1977.

Murphy, M. Lynne, *Semantic Relations and the Lexicon: Antonymy, Synonymy and Other Paradigms*, Cambridge University Press, 2003.

Nieuwerburgh, Stijn Van and Laura Veldkamp, "Learning Asymmetries in Real Business Cycles," *Journal of Monetary Economics*, 2006, *53* (4), 753–772.

Rozin, Paul and Edward B. Royzman, "Negativity Bias, Negativity Dominance, and Contagion," *Personality and Social Psychology Review*, 2001, *5* (4), 296–320.

Soroka, Stuart, Mark Daku, Dan Hiaeshutter-Rice, Lauren Guggenheim, and Josh Pasek, "Negativity and Positivity Biases in Economic News Coverage: Traditional Versus Social Media," *Communication Research*, 2018, *45* (7), 1078–1098.

Soroka, Stuart N., "Good News and Bad News: Asymmetric Responses to Economic Information," *The Journal of Politics*, 2006, *68* (2), 372–385.

Trier, Jost, *Der Deutsche Wortschatz im Sinnbezirk des Verstandes*, Heidelberg: Carl Winter, 1931.

Wei, Yanhao and Nikhil Malik, "Estimation Bias from Machine-Learned Features in Econometric Models," *Management Science*, 2026, *Accepted.*

Wierzbicka, Anna, *Semantics: Primes and Universals*, Oxford University Press, 1996.

Yang, Mochen, Edward McFowland, Gordon Burtch, and Gediminas Adomavicius, "Achieving Reliable Causal Inference with Data-Mined Variables: A Random Forest Approach to the Measurement Error Problem," *INFORMS Journal on Data Science*, 2022, *1* (2), 138–155.

_ , Gediminas Adomavicius, Gordon Burtch, and Yuqing Ren, "Mind the Gap: Accounting for Measurement Error and Misclassification in Variables Generated via Data Mining," *Information Systems Research*, 2018, *29* (1), 4–24.