

Measuring efficiency and equity framing in economics research: LLM-based evidence from 1950–2021

Measuring efficiency and equity framing in economics research

Sebastian Galiani^{1,*}, Ramiro H. Gálvez^{2,**}, Franco Mettola La Giglia³,
and Raul A. Sosa³

Abstract: We measure how frontier research frames what is normatively at stake along the efficiency and equity dimension. We develop and validate an LLM-based measurement pipeline and apply it to 27,464 full-text journal articles from 1950 to 2021. Efficiency focused framing rises through the late 1980s, then declines as equity related framing expands after 1990, especially in applied work and policy evaluations. By 2021, papers with an equity component are about as common as papers framed purely around efficiency. President transmittal letters in the Economic Report of the President show a similar post 1990 shift toward equity, providing an external benchmark.

Keywords: Normative Framing, Efficiency, Equity, LLMs, Text-as-Data

Classification: A14, B2, C8

*Correspondence address: sgaliani@umd.edu

**rgalvez@utdt.edu

1 Introduction

Economists have long analyzed public policy through the dual lenses of *efficiency* and *equity*, treating them as central and often competing normative objectives.¹ Because the two lenses emphasize different social outcomes, they are commonly viewed as involving trade-offs. Okun (1975) captured this idea in the “leaky bucket” metaphor: redistribution may advance equity, but it can also reduce efficiency when transfers dissipate resources or change behavior. This trade-off also invites an empirical question: have the relative weights placed on efficiency and equity shifted over time? Historical work on U.S. policy institutions suggests that, at least in policy discourse, they have. Berman (2022) argues that from the 1960s through the 1980s, policy evaluation increasingly adopted an “economic style” in which incentives, efficiency, and cost-benefit reasoning became central criteria for government action. What remains unclear is whether comparable shifts occur over longer horizons and whether they are confined to policy discourse.

This paper investigates that question. Our premise is that the contribution of research articles lies not only in the estimates they report or the models they develop, but also in how they frame what is at stake. Two papers can study the same economic object yet differ sharply in their normative framing. Contributions may be articulated in terms of optimality and aggregate welfare, in terms of distributional impacts and fairness, or in a way that is largely independent of both. Motivated by this distinction, we conduct a large-scale historical

¹In welfare analysis, efficiency concerns aggregate performance, namely the allocation of scarce resources so as to maximize total surplus or social welfare. Equity concerns the distribution of welfare, resources, or opportunities across individuals or groups, including inequality, poverty, and fairness. In our context, *normative framing* refers to how an argument motivates and interprets claims in these evaluative terms, meaning which objective is presented as being at stake.

analysis of frontier economics research to assess whether its normative framing has shifted systematically over time.

We analyze 27,464 full-text articles published from 1950 to 2021 in the so-called top-five general-interest economics journals, the *American Economic Review* (AER), *Econometrica* (ECTA), *Journal of Political Economy* (JPE), *Quarterly Journal of Economics* (QJE), and *Review of Economic Studies* (ReStud). For each article, we classify its normative framing. The classification is based on how papers motivate and interpret their contributions in welfare terms rather than on topic labels or keyword counts. To implement this classification at scale, we develop and validate an NLP strategy that combines predictions from frontier Large Language Models (LLMs) with more traditional text-based methods. The resulting procedure assigns each document to one of five mutually exclusive categories (*not a research article*, *other*, *efficiency*, *equity*, or *both*), yielding normative-framing labels for every article in our dataset.

In addition, we measure other corpus dimensions to characterize and benchmark our results. We use an NLP-based classifier to assign each article to one of four broad fields (*applied*, *applied theory*, *econometric methods*, and *theory*). Within applied work, we classify whether an article is a *policy evaluation*, defined as an empirical study whose central estimand is the causal effect of a concrete public intervention. Finally, we benchmark scholarly trends against policy rhetoric by constructing annual series from the President’s transmittal letters in the *Economic Report of the President* (ERP), using distinct measurement strategies to capture the relative emphasis on efficiency-oriented versus equity-oriented reasoning in high-level policy narratives throughout the sample period.

Our results point to a clear break between two periods. From 1950 through the late 1980s, efficiency-focused framing became increasingly prominent, rising to roughly 40% of documents in our corpus, while equity-related framing remained comparatively limited. Beginning in the early 1990s, the pattern shifted within research articles themselves. The efficiency share gradually receded toward about 30% by the end of the sample, while equity-related framing increased markedly. Papers classified as *equity* and *both* rose from roughly 10–15% before 1990 to about one-third of the corpus by 2021, so that papers with an explicit equity component became roughly as prevalent as purely efficiency-focused work. This post-1990 rise was concentrated in applied research and was closely linked to the growing prominence of policy-evaluation papers: applied papers published after 1990 incorporated equity considerations far more frequently than applied papers from earlier decades. We also find parallel movements in high-level policy rhetoric. Measures constructed from the ERP likewise shifted toward greater equity emphasis after 1990, mirroring the timing of the equity turn in economics research.

Our work connects to a growing methodological literature that treats text as data and develops tools to translate large corpora into measurable variables for studying economic and social phenomena (Grimmer and Stewart, 2013; Gentzkow et al., 2019; Ash and Hansen, 2023). Text-based measurement has been applied across domains that include media and political communication, finance, and policy analysis (see, for example, Tetlock, 2007; Gentzkow and Shapiro, 2010; Loughran and McDonald, 2011; Baker et al., 2016; Hansen et al., 2017). Early approaches often relied on dictionaries, word counts, and topic models. These methods, while effective for tracking salient terms and broad themes, can struggle when meaning is contextual, when the construct of interest is expressed indirectly, or when interpretation

depends on how an argument unfolds across sections of a document. Recent progress in deep learning has shifted attention toward contextual representations that better capture long-form text (Devlin et al., 2019; Dell, 2025). In particular, LLMs have enabled rubric-guided reading at the document level, which is useful when the object of interest is conceptual and implicit (Korinek, 2023; Ash et al., 2024; Bermejo et al., 2025). In parallel, we also speak to an emerging econometric literature that clarifies how to treat model outputs as measurement, emphasizing the importance of validation samples, benchmarking, and design safeguards when model-generated labels enter empirical analysis (Ludwig et al., 2025).

Our work also relates to literatures that motivate the focus on efficiency and equity and suggest that their relative salience can shift over time. Welfare economics distinguishes aggregate efficiency from distributive equity and emphasizes that welfare comparisons necessarily rest on normative judgments (Pareto, 1906; Pigou, 1920; Bergson, 1938; Samuelson, 1947; Arrow, 1951). Classic contributions also debated when efficiency and distribution can be separated in practice and what is lost when distributional judgments are set aside in applied evaluation (Kaldor, 1939; Hicks, 1939; Scitovsky, 1941; Robbins, 1932; Harsanyi, 1955). Building on these foundations, public economics operationalizes the efficiency–equity tension through social welfare functions, inequality measurement, and incentive-based analyses of taxation and redistribution (Atkinson, 1970; Sen, 1970; Mirrlees, 1971; Diamond and Mirrlees, 1971; Atkinson and Stiglitz, 1976; Saez, 2001), while Okun (1975) distilled the core intuition into a durable policy metaphor. In parallel, a large empirical literature documents the evolution of inequality and related distributional outcomes and has expanded the measurement of top incomes and the concentration of economic resources (Atkinson, 1997; Piketty and Saez, 2003; Piketty, 2014; Saez and Zucman, 2016). Related historical

work traces shifts in the language and evaluative criteria used in policy institutions (see, for example, Berman, 2022). Finally, we also speak to research that studies economics as a discipline by documenting long-run changes in methods and publication patterns, including the rise of design-based empirical work and shifts in what leading journals publish and reward (Ellison, 2002a,b; Angrist and Pischke, 2010; Card and DellaVigna, 2013; Hamermesh, 2013; Anauati et al., 2016; Angrist et al., 2017; Galiani et al., 2025).

This paper makes three contributions. First, it provides the first systematic long-run evidence on how frontier economics research frames contributions along the efficiency–equity dimension, documenting seven-decade patterns across journals and broad fields of study. Second, it develops an AI-based, rubric-guided measurement strategy for document-level analysis of large research corpora that produces auditable measures usable in empirical work at scale.² Third, it shows how AI-based measurement can be integrated into applied economics in a transparent and scientifically disciplined way, yielding interpretable variables from unstructured text that support replication, robustness checks, and downstream analysis.

The rest of the paper is structured as follows. Section 2 describes the data. Section 3 details the classification methodology. Section 4 presents the main results. Section 5 concludes.

²All code and input data implementing the measurement pipeline, as well as all variables used in the final analysis, will be made publicly available upon acceptance in the replication package.

2 Data

2.1 Full-text data on economics research articles

Consistent with prior work (Card and DellaVigna, 2013; Hamermesh, 2013; Anauati et al., 2016), we study long-run trends in economics using articles published in the so-called top-five journals. These outlets are a standard benchmark for the research frontier, reflecting a stable set of high-prestige journals whose editorial standards and disciplinary centrality are widely believed to shape what is produced and rewarded within the profession.

We obtain full-text data and document-level information from Constellate, a text analytics platform provided by ITHAKA, which also manages JSTOR and Portico. We query all documents published in AER, ECTA, JPE, QJE, and ReStud, obtaining machine-readable full text, bibliographic metadata, and platform-generated text statistics such as unigram counts. The raw download contains 39,214 documents, each tagged with a document subtype (“docSubType”). Of these, 27,464 are labeled “research-article”, 6,246 are labeled “book-review”, and 5,504 are labeled “misc.” We restrict the analysis to “research-article” documents. The Constellate export does not fully cover the last three years of the study window. In particular, JPE and QJE articles are not observed for 2019–2021, and ReStud articles are not observed for 2021.³

For descriptive context, Appendix Figure A1 summarizes the evolution of publication volume, average document length, and average authorship per article across journals over 1950–2021. Notably, even these raw descriptive patterns reproduce several well-documented

³All qualitative conclusions are unchanged when the analysis is truncated at 2018 (inclusive).

facts from the literature (see Card and DellaVigna, 2013), including the secular increase in article length and the rise in multi-authored papers.

2.2 Economic Report of the President transmittal letters

To benchmark scholarly trends against policy discourse, we construct a second text corpus based on the President’s annual transmittal letter to the U.S. Congress that opens the ERP. These letters summarize each administration’s assessment of economic conditions and policy objectives in language addressed to legislators and the broader public rather than to technical specialists, making them well suited for analyzing how efficiency and equity are articulated in high-level policy justification. We align this corpus with the journal analysis window (1950–2021), collecting one transmittal letter per year and treating each as a single annual observation.

3 Methodology

3.1 AI-based measurement of efficiency–equity framing in economics research articles

We develop an NLP measurement pipeline that captures the normative framing of economics research articles at scale. By *normative framing* we mean how a paper motivates and interprets its contribution in evaluative terms in the written text, namely the lens it uses to argue what is at stake. This is distinct from the authors’ personal normative commitments and from the welfare criterion assumed in the underlying model. Normative framing is also distinct from topical classification: two papers can study the same domain, such as public

health, yet differ in what they present as normatively relevant. One may emphasize aggregate welfare, surplus, productivity, or related notions of efficiency and optimality, for example by monetizing total benefits relative to costs. Another may emphasize distributional impacts, fairness, access, or inequality, for example by identifying who gains and who loses. Others may not engage the efficiency–equity axis, instead presenting descriptive facts, new methods, or theory without linking their contribution to efficiency or equity.

We implement a five-category labeling scheme that assigns each document to one mutually exclusive category. This yields a clean time series by filtering out non-original research material that remains in the corpus, such as notes or comments, and it distinguishes research articles by their normative framing along the efficiency–equity dimension. The five categories are

1. *Not a research article*: comments, corrigenda, editorials, replies, and related material that does not constitute original research.
2. *Other*: descriptive, methodological, or evaluative work that does not frame its contribution in terms of efficiency or distributive equity.
3. *Efficiency*: primary emphasis on efficiency, optimality, surplus, productivity, growth, or aggregate welfare, with limited and non-central distributional emphasis.
4. *Equity*: primary emphasis on inequality, poverty, fairness, access, opportunity, or redistribution, without a substantial efficiency or optimality framing.
5. *Both*: explicit attention to both dimensions, for example trade-offs, optimal redistribution, or interventions whose analysis foregrounds both aggregate and distributional consequences.

To assign each article a category, we use a two-stage procedure that combines zero-shot classifications from LLMs with a bag-of-words text classifier (see Jurafsky and Martin, 2009; Gentzkow et al., 2019).⁴ In stage one, multiple LLMs read the full text and apply the rubric below to produce initial labels. In stage two, a meta-classifier aggregates these outputs and combines them with additional text features to produce a single final label.

The rubric provided to the LLMs (see Appendix C.1 for a transcription) defines efficiency and equity in operational terms and lists the five mutually exclusive categories. It instructs the model to base its judgment on the paper’s stated research question and main findings, typically in the abstract, introduction, and conclusion, and to focus on the primary normative emphasis rather than incidental mentions. We feed the same rubric and the full text of each article to each LLM and obtain one classification per model.

Following Ludwig et al. (2025), we construct a validation set as a stratified sample of 1,059 articles by selecting up to three papers from each journal–year cell across the five journals. We then hand-label each article into one of the five categories using the same criteria provided to the LLMs, with labels assigned by author consensus. We use this labeled set to benchmark rubric-guided LLM performance and to train the meta-classifier that combines LLM predictions with standard NLP features.

We obtain zero-shot classifications from three LLMs for every document in the full corpus. The models are gpt-4o-mini-2024-07-18, gpt-5-mini-2025-08-07, and gpt-5-nano-2025-08-07, each able to process full-article inputs.⁵ Each model returns a structured JSON object with

⁴In NLP, zero-shot learning refers to a model’s ability to perform a task without task-specific training examples, relying instead on its general pre-trained knowledge and an instruction describing the task (Kojima et al., 2022).

⁵When providing an article to each model, we prepend basic bibliographic metadata, including the article title and journal name, to the article text. For gpt-4o-mini-2024-07-18 we set temperature to zero, and for the GPT-5 variants we request high reasoning effort.

a predicted category, a short justification, and a confidence rating. We use these outputs both to evaluate each base model against the 1,059 hand-labeled articles and as inputs to the supervised meta-classifier.⁶

The three models perform reasonably well overall, with variation across categories. The gpt-4o-mini-2024-07-18 model achieves 73% accuracy and the most balanced class performance (Appendix Table B1). The gpt-5-mini-2025-08-07 model achieves 71% (Appendix Table B3), and gpt-5-nano-2025-08-07 achieves 67% (Appendix Table B5). Across models, the non-research category is identified reliably. The main difficulty is separating *efficiency*, *equity*, and *both* (Appendix Tables B2, B4, and B6).

To improve performance, we fit a supervised meta-classifier that ensembles the zero-shot predictions (see Hastie et al., 2009). For each article in the hand-coded sample, we construct three feature blocks. First, we include each LLM predicted category as a categorical feature. Second, we add vote-count features recording how many models assign the article to each of the five categories.⁷ Third, we represent the text using unigram TF-IDF, which builds its vocabulary from the corpus itself.⁸ With these inputs, we estimate a regularized multinomial logistic regression with class balancing.⁹

The ensemble improves on any single zero-shot model and is our final classification procedure. In out-of-sample validation, it achieves 81% accuracy (Appendix Tables B7 and B8). Importantly, a TF-IDF-only classifier attains 66% accuracy, suggesting that this level of measurement precision is unlikely without recent AI advances and the use of LLMs as a

⁶For examples of hand-labeled articles and the corresponding gpt-5-mini-2025-08-07 explanations, see Appendix D.

⁷For example, if two models assign an article to *equity* and one assigns it to *both*, then the vote-count vector is (0, 0, 0, 2, 1) ordered as above.

⁸TF-IDF assigns higher weight to words that appear frequently within a document and lower weight to words that appear in many documents in the corpus, making it sensitive to historically distinctive terms. See Jurafsky and Martin (2009) for details.

⁹We select regularization and related hyperparameters using 10-fold cross-validation.

measurement technology. Appendix Figure A2 shows that out-of-sample predictions closely track the time-series patterns in the hand-coded data, indicating that the meta-classifier reliably recovers the long-run trends in normative framing.

We apply the meta-classifier to the full corpus. For articles in the hand-coded sample, we retain the human label. For all other articles, we assign the fitted meta-classifier prediction.

3.2 Measurement of field-of-economics classification

To characterize variation in normative framing across the discipline, we assign each article in our corpus to a broad field-of-economics-research category intended to capture the paper’s primary methodological orientation. Following Anauati et al. (2016), we assign each paper to exactly one of four mutually exclusive categories:¹⁰

1. *Applied* papers are empirically motivated and rely primarily on econometric or statistical analysis, or on descriptive statistics, as the core of the contribution. These papers are sometimes complemented by simple models that provide a theoretical framework.
2. *Applied theory* papers develop formal models to explain empirical phenomena, with empirical analysis playing a secondary role and typically limited to basic econometric exercises, simulations, or related techniques used to assess the implications of the model.
3. *Econometric methods* papers focus on the development of econometric or statistical methodology, including contributions related to data collection, identification strategies, data aggregation, or optimization techniques.

¹⁰Angrist et al. (2020) use a similar division but refer to these categories as “styles”.

4. *Theory* papers do not include an empirical fact section and primarily advance formal theoretical arguments, typically relying on mathematical analysis and logical reasoning, sometimes illustrated with numerical examples or simple calibrations.

Manually assigned field tags are available only for a subset of papers from prior work (11,624 articles), whereas our current corpus contains 27,464 articles. To obtain a complete set of field labels, we replicate the field-classification procedure in Galiani et al. (2025).¹¹ We start from the same labeled training sample used in that paper, which combines manually tagged articles from Anauati et al. (2016) and Anauati et al. (2020). Using this labeled sample, we train the same supervised text classifier as in Galiani et al. (2025). In brief, the method represents each article using a TF-IDF transformation of unigram features and estimates a regularized multinomial logistic regression with class reweighting (see, e.g., Jurafsky and Martin, 2009; Gentzkow et al., 2019; Kelly et al., 2021). This procedure achieves an out-of-sample accuracy of 87%, with particularly strong performance for applied, econometric methods, and theory, and comparatively weaker performance for applied theory. We then apply the trained classifier to the main corpus to predict a field tag for each of the 27,464 articles. We retain manual tags when available and use classifier-predicted tags otherwise.

Figure 1 traces the evolution of the field composition of publications in our corpus. In line with prior work, it documents a pronounced shift away from purely theoretical contributions toward applied and applied-theory research, with the reallocation accelerating in the late 1980s and especially during the 1990s.

¹¹ All details on sample construction, variable definitions, and the data construction workflow are provided in the replication package (Galiani et al., 2024).

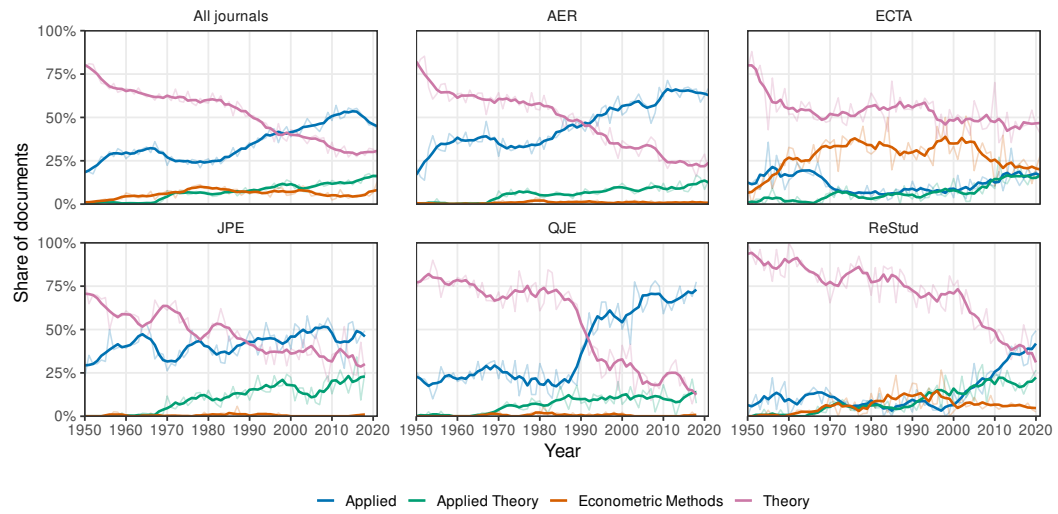


Figure 1. *Field composition of top five economics publications, 1950–2021.*

Notes. The figure reports annual shares of published documents in four mutually exclusive field categories *applied*, *applied theory*, *econometric methods*, and *theory*. Shares are shown for the pooled corpus and separately for each journal *American Economic Review* (AER), *Econometrica* (ECTA), *Journal of Political Economy* (JPE), *Quarterly Journal of Economics* (QJE), and *Review of Economic Studies* (ReStud). Solid lines show 5-year centered moving averages and faint lines show annual shares. Years with no observed documents are treated as missing and are not plotted.

3.3 AI-based measurement of policy-evaluation research within applied papers

To distinguish applied work that is organized around estimating the causal impact of a concrete public intervention (for example, a law, program, regulation, tax or subsidy, enforcement change, eligibility rule, or institutional reform) from applied work that is empirical but not centered on evaluating a specific policy, we construct a document-level indicator for whether an applied paper constitutes a *policy evaluation*.

We implement this indicator using the same two-stage procedure as for efficiency–equity framing (Section 3.1). In stage one, multiple LLMs apply a fixed rubric to each applied article in a zero-shot setting. The rubric (Appendix C.3) defines a policy evaluation as a study in which the *policy itself* is the object of interest and the estimand is its causal effect on outcomes, excluding cases where the policy mainly serves as a research-design device (e.g., an instrument or source of quasi-experimental variation to study a different relationship). In stage two, we aggregate LLM outputs using a meta-classifier trained on a manually labeled, year–journal stratified sample of 1,887 applied papers. The ensemble combines LLM predictions with standard text features, and we convert predicted probabilities into a binary label using a decision threshold chosen to maximize the out-of-fold F_1 score in stratified cross-validation. The resulting classifier achieves 95% out-of-sample accuracy (Appendix Tables B9 and B10) and closely reproduces the time-series patterns in the hand-labeled data (Appendix Figure A3).

Once trained, we apply the fitted classifier to the remaining applied papers to obtain the policy-evaluation indicator for the full corpus. For papers in the hand-labeled sample, we retain the manual label.

3.4 Multi-method measurement of efficiency–equity framing in policy narratives

To provide an external benchmark from policy discourse, we measure how ERP transmittal letters emphasize efficiency versus equity over time using three complementary indices.

First, we manually code each letter. We assign separate scores for efficiency and equity using a criterion that combines narrative emphasis and textual frequency. Each score ranges from 1 (low) to 3 (high). We summarize orientation with the equity score minus the efficiency score. This *manual-coded index* ranges from -2 to $+2$, with 0 indicating equal emphasis.

Second, we use a frontier LLM to score each letter along the efficiency–equity dimension. For each year, we provide the full text to GPT-5.2 with reasoning effort set to high, together with a fixed rubric that operationalizes efficiency and equity in policy discourse. The rubric in Appendix C.2 defines both concepts, uses the same three-point scale as the manual coding, and clarifies how to treat generic growth language, distributional rhetoric, and incidental mentions. We apply this procedure uniformly to all letters from 1950 through 2021. To address run-to-run variation, we run the rubric five times per letter-year, compute equity minus efficiency in each run, and use the average difference as the *LLM-coded index*.

Third, we construct a text-similarity measure that positions each letter by comparing its language to that of the 27,464 research articles whose normative framing we classify above. We concatenate each research article’s title, abstract, and full text, then use the resulting corpus to define a common vocabulary and TF–IDF weights. We represent both the research articles and each transmittal letter in this TF–IDF space. For each letter-year, we compute cosine similarity to every research article, rank articles by proximity, keep the top- k most similar articles (with $k = 50$), and build an annual index equal to the share of *equity* plus

both articles minus the share of *efficiency* articles in the top- k set. This *text-similarity index* infers orientation from proximity to articles with known framing rather than from direct scoring of the ERP text.

4 Results

4.1 Trends in normative framing in economics research

To characterize normative framing over time, Figure 2 plots, by publication year, the share of documents assigned to each of our five mutually exclusive categories defined in Section 3.1. The figure also reports *equity & both*, which aggregates articles classified as *equity* and *both* and captures the prevalence of papers with a substantive equity component, either as the primary emphasis or alongside efficiency.¹²

Two broad phases emerge. From 1950 through roughly the late 1980s, the share of articles classified as *efficiency* (top-left panel) rises steadily, from about one-fifth of published documents in the early postwar period to roughly two-fifths by the early 1980s, becoming the single predominant category. This rise is not matched by a decline in *equity* or *other*, which remain relatively stable. Instead, the main offsetting change is a sharp secular decline in documents classified as *not a research article* (bottom-right panel), from around 40% in the early decades to single digits by the 1990s.

From the early 1990s onward, the pattern shifts. The *efficiency* share stops rising and gradually declines, settling around 30% toward the end of the sample. Equity-related framing

¹²We show the full five-category decomposition because the *Not a research article* series is informative about long-run changes in journal content. As a robustness check, Appendix Figure A4 reproduces the series after dropping documents classified as *not a research article* and recomputing shares over research articles only. All qualitative patterns and conclusions are unchanged.

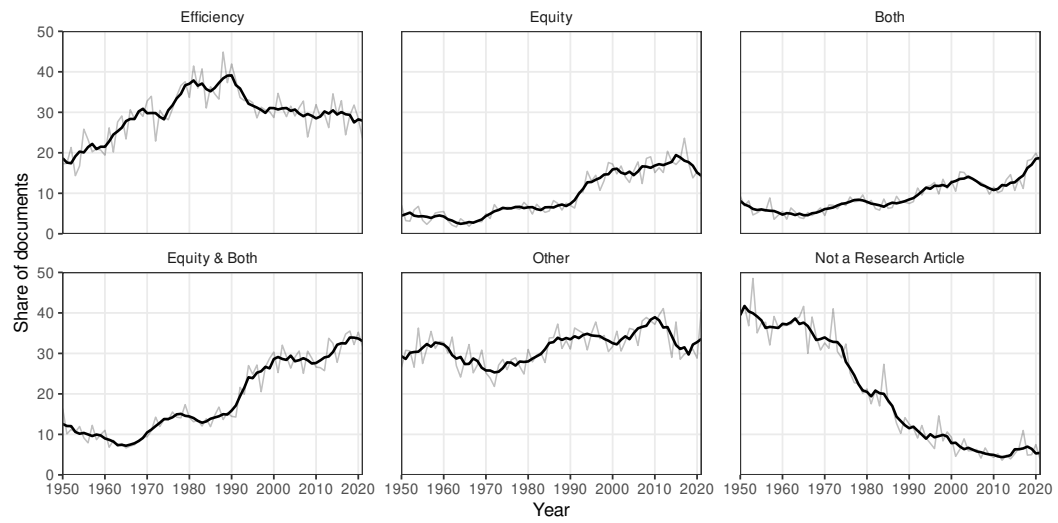


Figure 2. *Efficiency and equity framing in economics research, 1950–2021.*

Notes. The figure reports annual shares of documents assigned to five mutually exclusive categories *Not a research article*, *Other or no efficiency equity focus*, *Efficiency*, *Equity*, and *Both*. *Equity & Both* equals the sum of *Equity* and *Both*. Solid lines show 5-year centered moving averages and faint lines show annual shares. Years with no observed documents are treated as missing and are not plotted. Appendix Figure A4 reports the analogous trends computed over research articles only.

becomes more prominent. The share classified as *equity* increases sharply beginning in the 1990s (top-middle panel), and the share classified as *both* rises more gradually, especially after 2010 (top-right panel). These movements are summarized by *equity & both* (bottom-left panel), which increases from roughly 10–15% before 1990 to around 25–35% thereafter. By the end of the sample, papers with an explicit equity component are roughly as common as papers framed primarily around *efficiency*, indicating that distributional considerations have become a central element of normative framing.

Two additional features help interpret the trends. First, the *other* share is broadly stable over the full sample, which makes it unlikely that the post-1990 rise in equity-related framing mainly reflects papers being reclassified from method- or mechanism-driven work into normative categories. Second, because *not a research article* is already rare by the 1990s, the subsequent increase in *equity* and *both* must come from a reallocation within research articles themselves, not from further changes in the prevalence of non-research material.

4.2 Normative framing across fields of economics research

This subsection asks whether the aggregate shifts in Figure 2 are broad-based or concentrated in specific segments of the discipline. We first compare framing across broad fields of economics research (Section 4.2.1). We then focus on applied work and document the rising prominence and normative framing of policy evaluation (Section 4.2.2).

4.2.1 Efficiency and equity within fields

Figure 3 decomposes the aggregate trends in Figure 2 by broad field category. Each panel plots, within a field, the annual shares of research articles classified as *efficiency*, *equity*, and

equity & both. Shares are computed within field–year cells excluding documents classified as *Not a research article*.

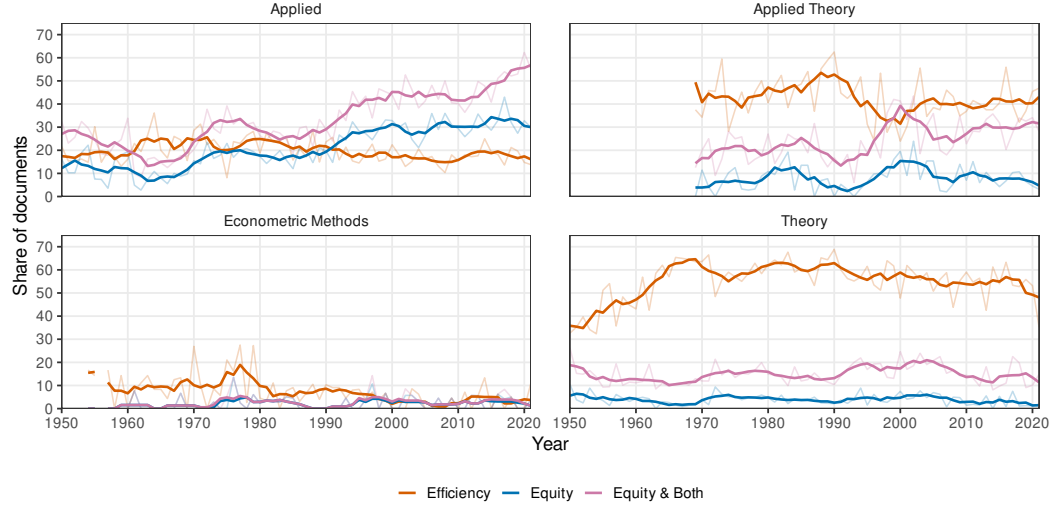


Figure 3. *Efficiency and equity framing within fields of economics research, 1950–2021.*

Notes. The figure reports, for each field category, annual shares of research articles classified as *efficiency focused*, *equity focused*, and *equity & both*. *equity & both* equals the sum of *equity focused* and *both efficiency and equity*. Shares are computed within field year cells and exclude documents classified as *not a research article*. Solid lines show 5-year centered moving averages and faint lines show annual shares. Field year cells with fewer than five articles are treated as missing and are not plotted.

Two patterns emerge. First, the long-run rise in equity-related framing is concentrated in applied work. Second, the broad fields differ sharply in how naturally their contributions map into the efficiency–equity dimension.

Within applied papers, equity began to gain substantial traction in the 1990s. The *equity* share rose steadily thereafter, and *equity & both* increased even more strongly, reflecting both an expansion of papers framed primarily around distributional outcomes and an increase in papers that integrated equity alongside efficiency. Notably, this shift did not displace efficiency framing. The *efficiency* share within applied work remained comparatively stable,

so applied economics became more distributionally attentive while continuing to produce a substantial volume of work motivated by aggregate efficiency.

Theory remained predominantly efficiency-oriented. The efficiency share rose through the 1950s and 1960s and then stayed high, while explicitly equity-centered theory remained scarce and *equity & both* showed virtually no post-1990 increase. Applied theory moved in the same direction but more mildly, with a 1990s reallocation away from purely efficiency framing toward *equity & both* as distributional considerations were more often integrated alongside efficiency or optimality arguments. Econometric methods papers were largely orthogonal to the efficiency–equity axis, with uniformly low shares in *efficiency*, *equity*, and *equity & both*. Most contributions instead fell into *other*, reflecting the field’s emphasis on identification, estimation, and measurement.

These within-field patterns help interpret the aggregate evolution in Figure 2. Figure 1 shows a large reweighting of the publication mix away from *theory* and toward *applied* and *applied theory*. Because equity-related framing grew most strongly within applied papers after 1990, this compositional shift amplified the aggregate rise of equity in the corpus. At the same time, the declining prominence of theory, where efficiency framing was most dominant and most stable, mechanically contributed to the decline in the aggregate *efficiency* share after the early 1990s. Overall, the post-1990 increase in equity-related framing reflected both within-field change, especially in applied research, and a shift in the publication mix toward fields in which equity considerations became more central.

4.2.2 Policy evaluation within applied research

We next assess whether the post-1990 rise in equity-related framing within applied work is linked to a particular type of paper by focusing on policy-evaluation research (see Section 3.3). Panel A of Figure 4 plots the share of applied papers classified as policy evaluations over time. Panel B shows how the efficiency–equity framing of policy-evaluation papers differs across two periods, 1950–1989 and 1990–2021.

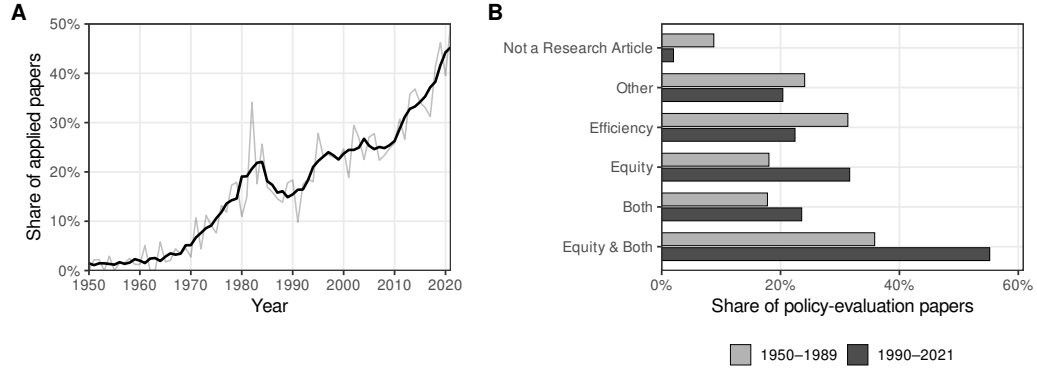


Figure 4. *Policy evaluation in applied research and normative framing.*

Notes. Panel A reports by year the share of *applied* papers classified as policy evaluations. Panel B reports, among policy evaluation papers, the shares in each efficiency/equity category in two periods 1950–1989 and 1990–2021, including *Equity & Both* defined as the sum of *Equity focused* and *Both efficiency and equity*. In Panel A, the dark line shows the 5-year centered moving average and the faint line shows annual shares.

Panel A shows that policy evaluation became an increasingly important component of applied research, with a pronounced acceleration from the 1990s onward and a large share of applied publications by the end of the sample. Panel B shows that the framing of policy evaluations shifted strongly toward equity after 1990. The *equity & both* share rose because both purely equity-framed evaluations and evaluations that integrated equity alongside efficiency became more common, while a substantial efficiency-oriented component remained.

Taken together, Figures 3 and 4 indicate that the post-1990 rise in equity-related framing coincided with two linked shifts. Applied work became a larger share of the corpus, and within applied work policy-evaluation papers became more common. Moreover, policy evaluation also became more equity-oriented after 1990, further contributing to the aggregate increase in equity-related framing.

4.3 Efficiency and equity framing in ERP transmittal letters

A natural question is whether the normative shift in economics research is also visible in policy discourse. We address this using the President’s ERP transmittal letters as an external benchmark. Figure 5 reports three annual series measuring each letter’s orientation along the equity–efficiency dimension, scaled so that higher values indicate stronger equity emphasis relative to efficiency.

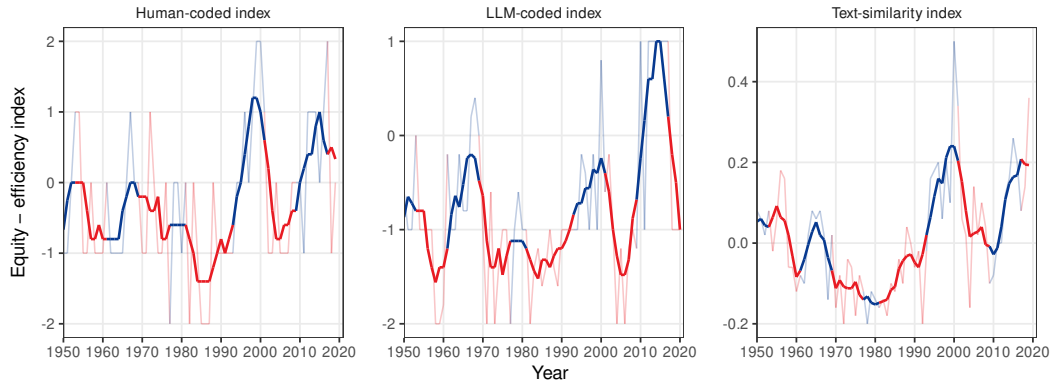


Figure 5. *Efficiency and equity emphasis in Economic Report of the President transmittal letters, 1950–2021.*

Notes. The figure reports three annual measures of transmittal letter orientation along the equity efficiency dimension. Higher values indicate greater equity emphasis relative to efficiency. Panels report the manual-coded index, the LLM-coded index, and the text-similarity index as defined in Section 3.4. Solid lines show 5-year centered moving averages and faint lines show annual shares. Line segments are colored by the party of the sitting President.

The three measures tell a broadly consistent story. From the 1950s through the 1980s, ERP narratives tended to place greater weight on efficiency-related than on equity-related themes, with substantial year-to-year variation, and the indices reached their lowest values in the 1980s. Beginning in the 1990s, the indices shifted upward, indicating rising equity emphasis in high-level policy discourse. The increase is especially pronounced in the manual and LLM series and is also visible in the text-similarity index (which has smaller amplitude by construction but moves in the same direction). The party coloring highlights an additional regularity. Equity emphasis is typically stronger under Democratic administrations and weaker under Republican administrations, although medium-run movements are not purely partisan and also reflect broader shifts that cut across administrations.

These policy-side trends align closely with the evolution of normative framing in economics research. In the journal corpus, efficiency-oriented framing rises through the late twentieth century, while equity-related framing becomes increasingly prominent after 1990. In the ERP, equity emphasis likewise becomes more pronounced after 1990. This parallel suggests that changes in academic framing and in the language used to justify economic policy moved in the same direction over the same period.

5 Discussion and conclusion

This paper studies how frontier economics research frames what is normatively at stake and how that framing evolved since 1950. We develop an AI-based, rubric-guided measure that we apply to nearly 28,000 economics research articles and benchmark against policy rhetoric in the ERP. We document a clear turning point in the time series. Efficiency-oriented framing

rose through the late 1980s, in part because non-research material occupied a shrinking share of journal space. After 1990, equity-related framing expanded sharply within research articles themselves. The combined prevalence of papers with an explicit equity component (*equity & both*) increased from roughly 10–15% before 1990 to around 35% by the end of the sample. This shift was concentrated in applied work and closely tracked the growing prominence of policy-evaluation papers. ERP transmittal letters exhibited a similar increase in equity emphasis over the same period.

Our analysis is descriptive, yet it establishes stylized facts that can be interpreted in light of broader methodological and institutional change. One plausible interpretation, offered as suggestive and not as causal evidence, is that the “credibility revolution,” richer microdata, and the diffusion of design-based empirical strategies increased the premium placed on questions that can be answered with transparent identification and defensible causal estimates. In this environment, it is relatively straightforward to document distributional consequences, including who gains and who loses and how impacts differ across groups. By contrast, strong efficiency claims typically require additional inputs, including a welfare metric, relevant counterfactuals, and general-equilibrium adjustments or spillovers. These ingredients often rest on stronger assumptions and harder-to-estimate parameters. As a result, it may have become more natural for papers to foreground distributional incidence rather than full welfare comparisons or optimal redistributive design.

Consistent with this interpretation, our data provide additional evidence that policy-evaluation research rarely translates estimated impacts into aggregate welfare accounting, even in settings where benefits and costs can be monetized. In such cases, a formal cost-benefit analysis could complement distributional incidence estimates by linking them to

explicit welfare comparisons. In a stratified sample of 953 policy evaluation papers by year and journal, we find that only 25, about 2.6%, contain a formal cost-benefit analysis.¹³ This pattern reinforces the idea that the expansion of equity-related framing within applied work has not typically been accompanied by a corresponding expansion of efficiency-style welfare accounting.

The ERP transmittal letters provide an external benchmark for the research-side patterns. When we measure efficiency and equity emphasis in these letters, we find a post-1990 rise in equity emphasis that mirrors the expansion of equity-related framing in economics research. We do not interpret this parallel as evidence that policy rhetoric drove academic framing, or the reverse. Instead, it makes it less likely that the research trend reflects only journal-specific dynamics, since a similar shift appears in a separate corpus written for a different audience and purpose. Importantly, the ERP benchmark is consistent with the possibility that changes in policy discourse and changes in empirical practice moved in parallel, with improved data and methods making distributional analysis easier to produce and defend, and the growing prominence of equity language making such evidence more central to how economists framed their contributions.

Overall, the evidence points to a discipline that has become more explicit about distributional stakes while relying less often on fully articulated welfare accounting. One implication is that the profession may benefit from linking the growing body of distributional evidence more directly to the policy-design questions emphasized in public economics, including how

¹³In this coding, we classify a paper as containing cost-benefit analysis only if it monetizes both benefits and costs and reports a formal summary comparison, for example net present value, internal rate of return, return on investment, or a benefit-cost ratio. Cost-effectiveness metrics such as cost per unit outcome and purely narrative discussions of costs and benefits do not qualify.

to redistribute and provide social insurance while accounting for behavioral responses, budget constraints, and equilibrium effects. Making that accounting more routine, even in partial and transparent form, would help bring the two lenses back together and would make empirical evidence on equity more directly usable for policy choice.

Affiliations

¹University of Maryland & NBER, USA

²Universidad Torcuato Di Tella, Argentina

³Universidad de San Andres, Argentina

References

- Anauati, M. V., Galiani, S., and Gálvez, R. H. (2020). Differences in citation patterns across journal tiers. *Economic Inquiry*, 58(3):1217–1232.
- Anauati, V., Galiani, S., and Gálvez, R. H. (2016). Quantifying the life cycle of scholarly articles across fields of economic research. *Economic Inquiry*, 54(2):1339–1355.
- Angrist, J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2017). Economic research evolves: Fields and styles. *American Economic Review*, 107(5):293–297.
- Angrist, J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2020). Inside job or deep impact? extramural citations and the influence of economic scholarship. *Journal of Economic Literature*, 58(1):3–52.
- Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30.
- Arrow, K. J. (1951). *Social Choice and Individual Values*. John Wiley & Sons, New York.
- Ash, E. and Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, 15(Volume 15, 2023):659–688.
- Ash, E., Hansen, S., and Muvdi, Y. (2024). Large language models in economics. CEPR Discussion Paper 19479, CEPR Press, Paris and London. Published 13 September 2024.
- Atkinson, A. and Stiglitz, J. (1976). The design of tax structure: Direct versus indirect taxation. *Journal of Public Economics*, 6(1):55–75.

- Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263.
- Atkinson, A. B. (1997). Bringing income distribution in from the cold. *The Economic Journal*, 107(441):297–321.
- Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty*. *The Quarterly Journal of Economics*, 131(4):1593–1636.
- Bergson, A. (1938). A reformulation of certain aspects of welfare economics. *The Quarterly Journal of Economics*, 52(2):310–334.
- Berman, E. P. (2022). *Thinking like an Economist: How Efficiency Replaced Equality in U.S. Public Policy*. Princeton University Press, Princeton, NJ.
- Bermejo, V. J., Gago, A., Gálvez, R. H., and Harari, N. (2025). Llms outperform outsourced human coders on complex textual analysis. *Scientific Reports*, 15(1):40122. Published 2025-11-17.
- Card, D. and DellaVigna, S. (2013). Nine facts about top journals in economics. *Journal of Economic Literature*, 51(1):144–161.
- Dell, M. (2025). Deep learning for economists. *Journal of Economic Literature*, 63(1):5–58.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

- Diamond, P. A. and Mirrlees, J. A. (1971). Optimal taxation and public production i: Production efficiency. *The American Economic Review*, 61(1):8–27.
- Ellison, G. (2002a). Evolving standards for academic publishing: A q-r theory. *Journal of Political Economy*, 110(5):994–1034.
- Ellison, G. (2002b). The slowdown of the economics publishing process. *Journal of Political Economy*, 110(5):947–993.
- Galiani, S., Gálvez, R. H., and Nachman, I. (2024). Ecin replication package for “Specialization Trends in Economics Research: A Large-Scale Study Using Natural Language Processing and Citation Analysis”. openICPSR deposit, Inter-university Consortium for Political and Social Research (ICPSR). Version V3. Ann Arbor, MI.
- Galiani, S., Gálvez, R. H., and Nachman, I. (2025). Specialization trends in economics research: A large-scale study using natural language processing and citation analysis. *Economic Inquiry*, 63(1):289–329.
- Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.
- Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71.
- Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.
- Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, 51(1):162–172.

- Hansen, S., McMahon, M., and Prat, A. (2017). Transparency and deliberation within the fomc: A computational linguistics approach*. *The Quarterly Journal of Economics*, 133(2):801–870.
- Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4):309–321.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY, 2 edition.
- Hicks, J. R. (1939). The foundations of welfare economics. *The Economic Journal*, 49(196):696–712.
- Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ, 2 edition.
- Kaldor, N. (1939). Welfare propositions of economics and interpersonal comparisons of utility. *The Economic Journal*, 49(195):549–552.
- Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–20.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

- Korinek, A. (2023). Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.
- Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.
- Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large language models: An applied econometric framework. Working Paper 33344, National Bureau of Economic Research.
- Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. *The Review of Economic Studies*, 38(2):175–208.
- Okun, A. M. (1975). *Equality and Efficiency: The Big Tradeoff*. The Brookings Institution, Washington, D.C.
- Pareto, V. (1906). *Manuale di economia politica con una introduzione alla scienza sociale*. Number 13 in Piccola Biblioteca Scientifica. Società Editrice Libreria, Milano.
- Pigou, A. C. (1920). *The Economics of Welfare*. Macmillan and Co., Limited, London.
- Piketty, T. (2014). *Capital in the Twenty-First Century*. Harvard University Press, Cambridge, MA.
- Piketty, T. and Saez, E. (2003). Income inequality in the United States, 1913–1998. *The Quarterly Journal of Economics*, 118(1):1–39.
- Robbins, L. (1932). *An Essay on the Nature and Significance of Economic Science*. Macmillan and Co., Limited, London.
- Saez, E. (2001). Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, 68(1):205–229.

- Saez, E. and Zucman, G. (2016). Wealth inequality in the united states since 1913: Evidence from capitalized income tax data. *The Quarterly Journal of Economics*, 131(2):519–578.
- Samuelson, P. A. (1947). *Foundations of Economic Analysis*. Harvard University Press, Cambridge, MA.
- Scitovsky, T. (1941). A note on welfare propositions in economics. *The Review of Economic Studies*, 9(1):77–88.
- Sen, A. K. (1970). *Collective Choice and Social Welfare*. Number 5 in Mathematical Economics Texts. Holden-Day, San Francisco, Calif.
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Online Appendix to:

Measuring efficiency and equity framing in economics
research: LLM-based evidence from 1950–2021

Sebastian Galiani^{1,2}, Ramiro H. Gálvez³, Franco Mettola La Giglia⁴, Raul A. Sosa⁴

¹University of Maryland, ²NBER, ³Universidad Torcuato Di Tella, ⁴Universidad de San Andres

A Appendix figures

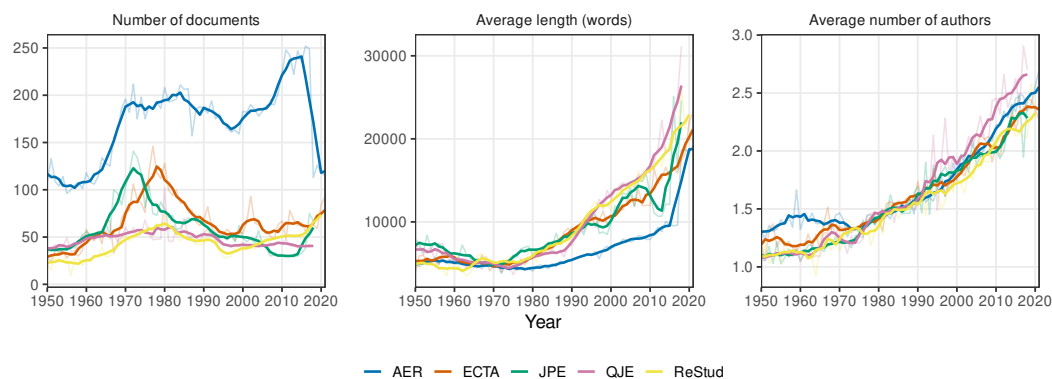


Figure A1. *Trends in the top five journal corpus, 1950–2021.*

Notes. The left panel shows the annual number of published documents. The middle panel shows average document length in words. The right panel shows the average number of authors per document. Lines correspond to the *American Economic Review* (AER), *Econometrica* (ECTA), *Journal of Political Economy* (JPE), *Quarterly Journal of Economics* (QJE), and *Review of Economic Studies* (ReStud).

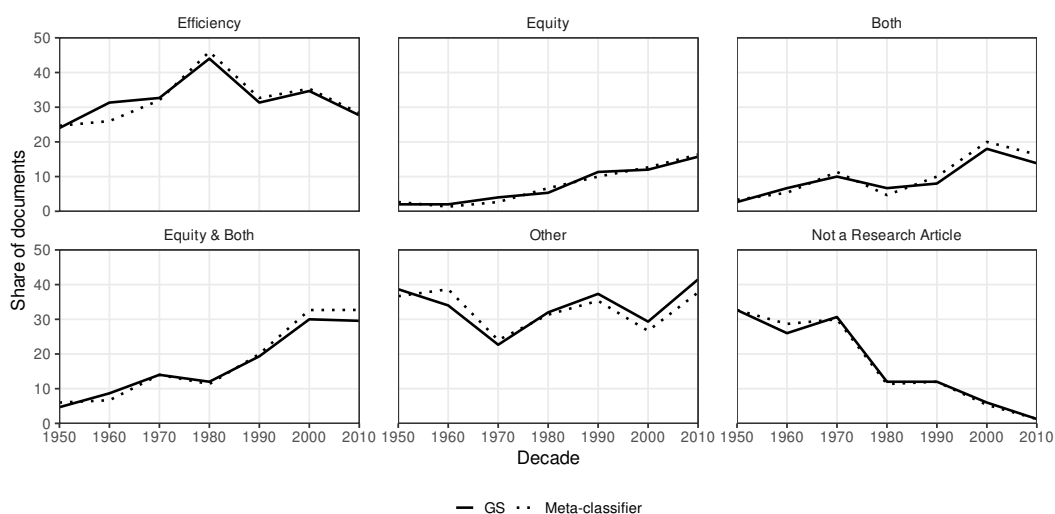


Figure A2. *Validation of category trend replication using gold standard labels and the meta classifier, 1950–2021.*

Notes. The figure compares decade level shares of documents across efficiency equity categories for two label sources. Solid lines with markers show gold standard labels obtained as out of fold predictions from stratified ten fold cross validation on the hand coded sample. Dotted lines show the corresponding meta classifier predictions. The sample is restricted to documents in the hand coded set for which both series are available. Documents from 2020–2021 are pooled with the 2010s so the final decade combines the 2010s with 2020–2021.

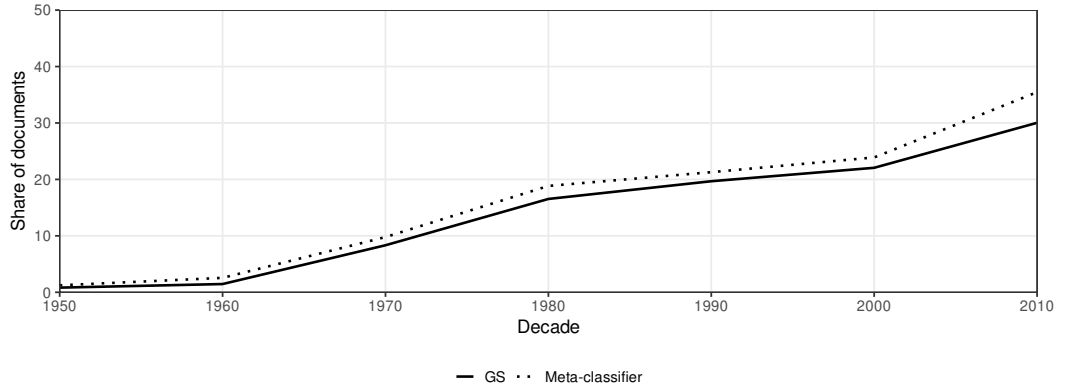


Figure A3. *Validation of policy evaluation trend replication using gold standard labels and the meta classifier, 1950–2021.*

Notes. The figure compares decade level shares of documents classified as policy evaluations for two label sources. Solid lines show gold standard labels obtained as out of fold predictions from stratified ten fold cross validation on the hand coded sample. Dotted lines show the corresponding meta classifier predictions. The sample is restricted to documents in the hand coded set for which both series are available. Documents from 2020–2021 are pooled with the 2010s so the final decade combines the 2010s with 2020–2021.

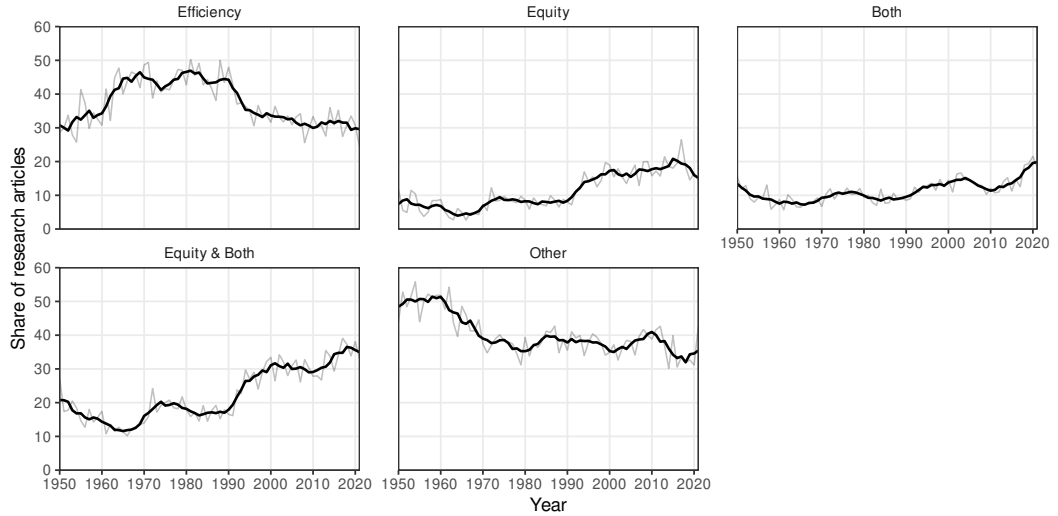


Figure A4. *Efficiency and equity framing in economics research using research articles only, 1950–2021.*

Notes. The figure reports annual shares of research articles only and excludes documents classified as *Not a research article*. Shares are shown for four mutually exclusive categories *Other or no efficiency equity focus*, *Efficiency*, *Equity*, and *Both*. *Equity & Both* equals the sum of *Equity* and *Both*. Solid lines show 5-year centered moving averages and faint lines show annual shares.

B Appendix tables

Table B1. *Validation predictive performance for GPT-4o-mini-2024-07-18*

	Precision	Recall	F_1 -Score	n
Not a research article	0.93	0.90	0.91	181
Other	0.73	0.83	0.77	357
Efficiency	0.86	0.55	0.67	341
Equity	0.69	0.62	0.66	80
Both	0.42	0.79	0.55	100
All (weighted average)	0.77	0.73	0.73	1059

Notes. Precision is equal to the proportion of all observations predicted as belonging to a given class that were correctly classified. Recall is equal to the proportion of all observations effectively belonging to a given class that were correctly classified. The F_1 -Score is equal to the harmonic mean between precision and recall.

Table B2. *Confusion matrix for GPT-4o-mini-2024-07-18*

	Not a research article	Other	Efficiency	Equity	Both
Not a research article	162	12	4	2	1
Other	11	296	20	9	21
Efficiency	1	86	186	3	65
Equity	0	6	1	50	23
Both	0	7	6	8	79

Notes. Rows represent the actual class and columns represent the predicted class.

Table B3. *Validation predictive performance for GPT-5-mini-2025-08-07*

	Precision	Recall	F_1 -Score	n
Not a research article	0.85	0.90	0.87	181
Other	0.77	0.65	0.70	357
Efficiency	0.65	0.81	0.73	341
Equity	0.59	0.51	0.55	80
Both	0.54	0.41	0.47	100
All (weighted average)	0.71	0.71	0.71	1059

Notes. Precision is equal to the proportion of all observations predicted as belonging to a given class that were correctly classified. Recall is equal to the proportion of all observations effectively belonging to a given class that were correctly classified. The F_1 -Score is equal to the harmonic mean between precision and recall.

Table B4. *Confusion matrix for GPT-5-mini-2025-08-07*

	Not a research article	Other	Efficiency	Equity	Both
Not a research article	162	10	5	2	2
Other	19	231	93	5	9
Efficiency	7	36	277	9	12
Equity	2	18	7	41	12
Both	0	6	41	12	41

Notes. Rows represent the actual class and columns represent the predicted class.

Table B5. *Validation predictive performance for GPT-5-nano-2025-08-07*

	Precision	Recall	F_1 -Score	n
Not a research article	0.90	0.82	0.86	181
Other	0.84	0.49	0.62	357
Efficiency	0.57	0.87	0.69	341
Equity	0.76	0.35	0.48	80
Both	0.46	0.58	0.51	100
All (weighted average)	0.72	0.67	0.66	1059

Notes. Precision is equal to the proportion of all observations predicted as belonging to a given class that were correctly classified. Recall is equal to the proportion of all observations effectively belonging to a given class that were correctly classified. The F_1 -Score is equal to the harmonic mean between precision and recall.

Table B6. *Confusion matrix for GPT-5-nano-2025-08-07*

	Not a research article	Other	Efficiency	Equity	Both
Not a research article	149	11	15	1	5
Other	14	174	148	5	16
Efficiency	2	19	297	1	22
Equity	0	2	24	28	26
Both	0	0	40	2	58

Notes. Rows represent the actual class and columns represent the predicted class.

Table B7. *Validation predictive performance for Meta classifier*

	Precision	Recall	F_1 -Score	n
Not a research article	0.93	0.94	0.94	181
Other	0.81	0.79	0.80	357
Efficiency	0.82	0.82	0.82	341
Equity	0.70	0.70	0.70	80
Both	0.62	0.67	0.64	100
All (weighted average)	0.81	0.81	0.81	1059

Notes. Precision is equal to the proportion of all observations predicted as belonging to a given class that were correctly classified. Recall is equal to the proportion of all observations effectively belonging to a given class that were correctly classified. The F_1 -Score is equal to the harmonic mean between precision and recall.

Table B8. *Confusion matrix for Meta classifier*

	Not a research article	Other	Efficiency	Equity	Both
Not a research article	170	7	3	1	0
Other	11	283	40	13	10
Efficiency	1	46	278	2	14
Equity	0	7	0	56	17
Both	0	6	19	8	67

Notes. Rows represent the actual class and columns represent the predicted class.

Table B9. *Validation predictive performance for meta-classifier*

	Precision	Recall	F_1 -Score	n
Other in applied	0.99	0.96	0.97	1612
Policy evaluation	0.80	0.92	0.85	275
All (weighted average)	0.96	0.95	0.96	1887

Notes. Precision is equal to the proportion of all observations predicted as belonging to a given class that were correctly classified. Recall is equal to the proportion of all observations effectively belonging to a given class that were correctly classified. The F_1 -Score is equal to the harmonic mean between precision and recall.

Table B10. *Confusion matrix for meta-classifier*

	Other in applied	Policy evaluation
Other in applied	1549	63
Policy evaluation	23	252

Notes. Rows represent the actual class and columns represent the predicted class.

C Instructions provided to the LLMs

This appendix documents the instruction blocks provided to the LLMs for each measurement task in the paper. The instructions specify the task definition and response format, and they operationalize the coding criteria applied to the underlying text.

C.1 Research articles’ normative-framing classification

For the article-level classification of normative framing, the LLM received the following rubric and required output format, followed by the full text of the article.

Below you are provided with authoritative instructions for classifying articles, followed by an article to classify. Read the instructions carefully and treat them as binding guidelines. Then read the article and return your classification exactly as requested.

***Instructions*:**

Your task is to assign the document to the correct category with high accuracy, strictly following the definitions and guidelines below. Apply the rules consistently and without exception.

***Definitions*:**

Efficiency refers to the optimal use of scarce resources to maximize output or welfare at the aggregate level. It concerns optimal allocation, productivity, total surplus, economic growth, and cost effectiveness. In economics, efficiency asks whether resources are arranged and used in a way that improves overall performance or welfare.

Equity refers to fairness in the distribution of resources, opportunities, and outcomes across individuals or groups. It concerns distributional outcomes such as inequality, poverty, mobility, and discrimination, as well as the well being of specific populations. In economics, equity asks who gains and who loses and whether distributions are fair.

***Classification categories*:**

1. Not a Research Article: Use this when the document is not a complete academic research article. Typical cases include editorials, comments, replies, notes, essays, corrigenda, errata, book reviews, article reviews, introductions, or conference sections. These types may be identified in the title, abstract, or opening paragraph. Any document that does not present original research findings belongs here. All notes, comments, and replies must be classified in this category regardless of their content, whether theoretical, empirical, or of any other type. If note, comment or reply is mentioned in the title, the article must be classified as 1.

2. Research Not Focused on Efficiency or Equity: Use this when it is a research article but neither efficiency nor equity is a central analytical objective. Typical cases include descriptive work, papers primarily focused on measurement or developing new measurement techniques, methods papers, and theory or empirics that study mechanisms or causality without framing results in terms of optimality or distributive justice.

When evaluating whether a paper belongs here, focus on the nature and implications of the outcomes studied, not just the methodology or presence of causal inference. Papers that present models or fit theoretical models to data without evaluating policy or welfare implications belong here. Program

evaluations that only test whether something works, without an efficiency or equity lens, also belong here.

Do not default a paper to this category just because it is labeled a survey or review.

3. Efficiency Focused Research: Use this when the article's primary aim is to analyze optimal allocation, aggregate welfare, productivity, or cost-benefit. The efficiency lens must guide the main question and the conclusions. Require explicit references to optimality, efficiency, or welfare maximization.

References to Pareto efficiency are a strong signal of an efficiency focus.

The following do not by themselves indicate an efficiency focus: mere mentions of outcomes, effects, results, or "effective" policies; growth (unless framed in terms of optimal resource use); program effectiveness; the presence of a model, production function, or general equilibrium setup. Efficiency in this context refers to economic efficiency---typically concerning policy choices, resource allocation decisions, or institutional design---not statistical or methodological efficiency.

Pay attention to whether the variables under study are decision variables: if the paper focuses on the role of non-decision variables (such as shocks to expectations or exogenous factors), it is less likely to be about efficiency. Papers that study markets with distortions or inefficiencies (such as information asymmetries) do not necessarily have an efficiency focus unless they evaluate solutions or optimal responses to these distortions.

Classify as efficiency only when the evaluation is framed around optimal resource use or aggregate welfare as the main criterion.

4. Equity Focused Research: Use this when the article's primary aim is to analyze distributional outcomes or fairness. Inequality, poverty, mobility, discrimination, or group-specific welfare must guide the main question and the conclusions.

Group-specific welfare includes outcomes such as wage differentials across groups, female labour participation, or the well-being of vulnerable populations. Papers that discuss social or redistributive aspects should be tagged with equity (either category 4 or 5), even if they do not use explicit normative language. Papers that report distributional facts or obtain distributional consequences as their main findings belong here.

Implicit equity language is acceptable when the conclusions hinge on who gains and who loses. Do not require explicit normative terms if the distributional lens is clearly central in the main sections.

However, if distributional concerns are only mentioned in passing or are secondary to the main research question, the paper does not belong here. Topics like governance are not automatically equity unless they focus on distributional outcomes.

5. Both Efficiency and Equity: Use this category when the article explicitly addresses both lenses. The paper analyzes aggregate performance and distributional outcomes in a balanced way. Do not reserve this only for perfectly symmetric treatments. Apply it whenever both goals are clearly integral to the research question and neither is incidental. However, be careful to distinguish primary from secondary concerns: the presence of heterogeneity in general equilibrium models does not automatically indicate an equity focus, and brief mentions that "some group is worse off" in an efficiency-focused paper do

not qualify unless distributional concerns are central to the research hypothesis. Both efficiency and equity considerations must be substantial parts of the analysis, not just passing observations.

***Additional guidelines*:**

Base your judgment exclusively on the abstract, introduction, and conclusion, as these sections state the research question, objectives, and main findings of the paper. Focus particularly on the implications and consequences presented in the results and conclusions---what the paper finds matters more than the methods used.

Distinguish carefully between the paper's primary focus and secondary or incidental mentions. The stated research objective and the nature of the main findings determine the classification, not the mere presence of technical elements like models, heterogeneity, or distributional data. A paper's category is defined by what it aims to accomplish and what it concludes, not by the tools it employs along the way.

Ignore all footnotes and endnotes. You must accurately recognize both explicit and implicit references to the concepts of efficiency and equity (which we define below), even when these ideas are conveyed indirectly, through diverse language, or without using the exact terms.

***Response format*:**

Return the final answer in plain JSON with exactly the following keys:

- "category": an integer from 1 to 5 corresponding to the assigned category

- ‘‘short_explanation’’: a brief explanation (no more than 4 sentences) justifying the classification, referencing the presence or absence of efficiency/equity considerations
- ‘‘confidence’’: an integer from 1 to 3 that indicates the model’s confidence level in the classification. Use 1 for high confidence (the document clearly fits one category based on its keywords, arguments, and focus), 2 for medium confidence (the document includes elements from more than one category or has some ambiguity, but leans toward one), and 3 for low confidence (the document is difficult to classify due to unclear language, mixed content, or lack of identifiable indicators)

The JSON must be valid and contain only these keys. Each key should be present just once. Do not include Markdown, headers, comments, code blocks, or additional text of any kind.

***Article*:**

C.2 Economic Report of the President transmittal-letter coding

For the analysis of Economic Report of the President (ERP) transmittal letters, the LLM received the following coding instructions prior to each letter. These instructions define how to map the text into structured outputs capturing its emphasis on efficiency and equity.

Below, you will receive the introductory message to the yearly Economic Report of the President, addressed to the Congress of the United States.

***Instructions*:**

Your task is to assess how much the text addresses economic efficiency and how much it addresses economic equity, by assigning two separate scores. Determine these scores using a single criterion that combines narrative emphasis and thematic recurrence. Place greatest weight on the introduction and the primary framing passages, where administrations typically state their perspective and articulate the normative priorities that organize the policy program, and also consider how frequently the letter returns to efficiency coded and equity coded language throughout the document.

1. efficiency_score (1--3): intensity of efficiency-related themes

Definition (what ‘‘efficiency’’ means here):

Efficiency emphasis is about improving how effectively the economy turns resources into output and well-being | e.g., productivity, innovation, competition, allocative efficiency, incentive effects, reducing distortions, and policies framed around raising long-run productive capacity.

Important clarification:

General references to ‘‘growth,’’ ‘‘a strong economy,’’ ‘‘job creation,’’ or ‘‘lower inflation,’’ or ‘‘sound fiscal management’’ do not by themselves indicate a strong emphasis on efficiency. These themes are more appropriately counted toward efficiency when the text explicitly connects them to productivity, incentives, allocative efficiency, competition, or long-run productive capacity, and when such connections receive more than passing attention.

Scoring anchors:

* 1 = low emphasis on efficiency

- Efficiency-related ideas are absent or only incidental.

- Mentions are brief, generic, or purely descriptive (e.g., reporting economic conditions without an efficiency-oriented policy frame).

* 2 = moderate emphasis on efficiency

- The text contains meaningful efficiency-oriented content (e.g., discussion of productivity, competitiveness, innovation, incentives, market functioning, reducing distortions), but it is one theme among several.

- Efficiency is discussed substantively at least once or appears repeatedly but does not clearly organize the overall message.

* 3 = strong emphasis on efficiency

- Efficiency/productivity/incentive/allocative themes are a central organizing frame of the introduction.

- The text returns to these themes multiple times and uses them to structure the main economic agenda (e.g., a sustained emphasis on raising productivity, improving incentives, expanding long-run capacity, or major efficiency-oriented reforms).

Examples of efficiency-related themes (when emphasized substantively):

- productivity and potential output; competitiveness; innovation and R&D; investment incentives

- improving resource allocation; competition and market functioning; reducing distortions

- deregulation or regulatory reform framed around efficiency

- tax reform framed around incentives/efficiency (not merely revenue levels)

- trade liberalization framed around efficiency/competition/productivity

- deficit reduction framed as supporting private investment and capital formation by reducing ‘‘crowding out’’
- fiscal or monetary policy framed explicitly as supporting long-run productive capacity or efficient allocation

2. equity_score (1--3): intensity of equity- or fairness-related themes

Definition (what ‘‘equity’’ means here):

Equity emphasis is about distribution, fairness, inclusion, equal opportunity, and economic security|i.e., who gains/loses from economic outcomes and policies, and whether benefits are broad-based or targeted toward disadvantaged groups.

Important clarification:

Equity emphasis may be expressed without using the word ‘‘equity.’’ Count explicit distributional language (poverty/inequality/fairness) and also clear ‘‘shared prosperity/broad-based growth,’’ ‘‘equal opportunity,’’ inclusion framing|especially when paired with concrete policies or sustained discussion.

Additional clarification:

Many introductions include brief, noncentral ‘‘compassion’’ or ‘‘safety net’’ language (e.g., ‘‘protect the truly needy,’’ ‘‘opportunity for all,’’ ‘‘help families,’’ ‘‘no one left behind’’) without sustained discussion of distribution or specific groups/policies. Treat this as weak equity evidence. Unless the introduction devotes a substantive portion of the text (e.g., a full paragraph or repeated discussion) to distributional outcomes, disadvantaged groups, or concrete protective/redistributive policies, the equity_score will usually remain 1.

Scoring anchors:

* 1 = low emphasis on equity

- Equity/fairness/distributional themes are absent or only token.
- At most, there are vague statements like ‘‘for all Americans’’ without substantive distributional framing or specific groups/policies.

* 2 = moderate emphasis on equity

- The text gives meaningful attention to distributional outcomes, fairness, opportunity, or economic security, but this is one theme among several.
- Equity is discussed substantively more than once (e.g., repeated distributional framing and/or multiple concrete policies aimed at specific groups). A single reassurance sentence or one isolated program mention is generally not enough for a 2.
- Examples include: sustained discussion of middle-class/working-family outcomes; poverty or hardship; equal opportunity; access to education/health; or specific policies aimed at protecting/assisting particular groups.

* 3 = strong emphasis on equity

- Equity/fairness/distributional concerns are a central organizing frame of the introduction.
- The text repeatedly emphasizes who benefits, who is being protected or lifted up, or how growth/policy should be shared; it devotes multiple substantive passages to these concerns and/or centers the agenda around them.

Examples of equity-related themes (when emphasized substantively):

- poverty, inequality, distribution of gains/losses; fairness; ‘‘those left behind’’
- middle-class living standards; wages; job quality; labor protections

- disadvantaged groups (low-income households, unemployed, minorities, distressed regions, rural/urban hardship)
- social insurance/safety nets (health coverage, retirement security, unemployment insurance, food assistance, housing support)
- equal opportunity, inclusion, anti-discrimination, access to education/health/childcare
- ‘‘shared prosperity’’ or ‘‘broad-based growth’’ framing when developed beyond a slogan

***General note*:** Use the full 1-3 scale for both efficiency and equity. Score each based on its relative prominence in the text, and do not hesitate to assign a 1 when a theme is clearly much less important than the other. Reserve a score of 2 for cases where efficiency or equity/fairness is genuinely one of several major themes, rather than clearly dominant or clearly minor.

***Output format*:**

For each report, output only one JSON object, with the following keys:

*** efficiency_score** (integer 1-3)

*** equity_score** (integer 1-3)

*** explanation** (a short paragraph explaining the rationale for each score)

The JSON must be valid and contain only these keys. Each key should be present just once. Do not include Markdown, headers, comments, code blocks, or additional text of any kind.

***Report to analyze*:**

C.3 Policy-evaluation classification

To determine whether a paper constitutes a policy evaluation under our definition, the LLM was provided the following instruction block and response specification, followed by the article text.

Below you are provided with authoritative instructions for classifying articles, followed by an article to classify. Read the instructions carefully and treat them as binding guidelines. Then read the article and return your classification exactly as requested.

***Instructions*:**

Your task is to assign the document to the correct category with high accuracy, strictly following the definition and guidelines below. Apply the rules consistently and without exception.

***Definition*:**

Policy evaluation is the empirical assessment of what a concrete public intervention does. A study qualifies as policy evaluation when the policy itself is the object of interest and the analysis is organized around estimating what impact policy P has on outcome Y. In applied economics, this means measuring the causal effect of a specific policy, such as a law, program, regulation, tax or subsidy, enforcement change, eligibility rule, or institutional reform, on well defined outcomes relative to a credible counterfactual.

Methodologically, policy evaluation relies on empirical designs that exploit plausibly exogenous variation in policy exposure to identify causal effects. Common implementations include ordinary least squares regressions with controls and fixed effects, instrumental variables, difference in differences and event

study designs, regression discontinuity and related threshold or kink designs, randomized controlled trials and encouragement designs, synthetic control methods, and matching or reweighting approaches. The resulting estimates capture the total effect of the policy as implemented, rather than parameters from a fully specified structural model that separately models underlying mechanisms.

Clarification: if a policy is used mainly as an instrument or source of variation to study a different causal relationship, rather than to estimate the policy's own effect on outcomes, the study is not considered a policy evaluation.

***Classification categories*:**

* 1 = Policy evaluation: The paper meets the definition of policy evaluation.

* 0 = Not policy evaluation: The paper does not meet the definition of policy evaluation.

***Response format*:**

Return the final answer in plain JSON with exactly the following keys:

- "category": an integer equal to 1 or 0 (1 = Policy evaluation, 0 = Not policy evaluation), following the category definitions above.

- "short_explanation": a brief explanation (no more than 4 sentences) justifying whether the paper is classified as policy evaluation or not.

- "confidence": an integer from 1 to 3 indicating confidence in the classification. Use 1 for high confidence (clearly policy evaluation or clearly not), 2 for medium confidence (some ambiguity but leans one way), and 3 for low confidence (unclear/mixed, hard to determine from the text).

The JSON must be valid and contain only these keys. Each key should be present just once. Do not include Markdown, headers, comments, code blocks, or additional text of any kind.

***Article*:**

D Examples of rubric-guided coding decisions

This appendix section illustrates the application of the guidelines in Appendix C. For each guideline, we present representative examples covering the relevant categories or score anchors and briefly justify the assigned labels using the decision rules specified in the instructions.

D.1 Normative-framing classification

Following Appendix C.1, each document is assigned to exactly one of five mutually exclusive categories. The examples below are intended to make the rubric operational: for each category, we provide one representative paper, a brief justification in the spirit of the rubric (i.e., emphasizing how the paper frames its core contribution), and then the corresponding LLM rationale recorded at the time of coding.

Category 1: Not a research article. Elliott, James R. (1961). *“The Burden of the Public Debt: Comment.”*

This document is a *comment* rather than a standalone research article and therefore does not qualify for inclusion as a research article under the rubric.

LLM explanation (gpt-5-mini-2025-08-07): The document is explicitly a short comment and a reply (title: ‘Comment’ and contains a ‘Reply’) published as brief communications, not a complete original research article. It critiques and discusses another paper rather than presenting new empirical or theoretical research, and thus does not itself frame a primary efficiency or equity analysis.

Category 2: Research not focused on efficiency or equity. Hausman, Jerry A., and William E. Taylor. (1983). *“Identification in Linear Simultaneous Equations Models with Covariance Restrictions: An Instrumental Variables Interpretation.”*

The paper is a methodological contribution in econometrics, focusing on identification and estimation in simultaneous-equation systems, and it does not motivate its contribution in terms of welfare optimality or distributional fairness; any references to “efficiency” are in the statistical (estimator) sense, which the rubric treats as orthogonal to the efficiency–equity axis.

LLM explanation (gpt-5-mini-2025-08-07): This is a theoretical econometrics research article deriving identification and estimation conditions (IV, 2SLS, FIML) for linear simultaneous-equation models with covariance restrictions. Its primary focus is methodological—identifiability and instrument construction—not on economic efficiency in the welfare/optimal allocation sense or on distributional/equity outcomes. While it mentions estimator (statistical) efficiency, it does not analyze aggregate welfare or distributional fairness.

Category 3: Efficiency-focused research. Kihlstrom, Richard E., and Jean-Jacques Laffont. (1983). *“Implicit Labor Contracts and Free Entry.”*

The paper is organized around welfare and efficiency concepts, explicitly developing and

comparing first-best and second-best efficiency criteria across alternative market institutions, so that the central evaluative standard is aggregate optimality rather than distributive incidence.

LLM explanation (gpt-5-mini-2025-08-07): The paper’s primary aim is to analyze optimality and welfare properties of equilibria: it explicitly defines and studies Arrow–Debreu (first-best) and several second-best efficiency concepts (contractual, Diamond, sure-wage, spot-wage), proves existence/uniqueness results and Pareto comparisons across market institutions. Distributional issues (who becomes entrepreneur vs worker and risk-aversion sorting) are discussed but are secondary to the central efficiency/welfare analysis.

Category 4: Equity-focused research. Rubinstein, Yona, and Dror Brenner. (2014). *“Pride and Prejudice: Using Ethnic-Sounding Names and Inter-Ethnic Marriages to Identify Labour Market Discrimination.”*

The paper’s core contribution is a distributional one: it documents and interprets differential labor-market outcomes across ethnic groups (wage penalties tied to perceived ethnicity), placing “who is harmed” at the center rather than framing results around aggregate welfare maximization.

LLM explanation (gpt-5-mini-2025-08-07): The paper’s primary aim is to analyze distributional outcomes and fairness: it documents ethnic discrimination in wages between Sephardic and Ashkenazi groups using surnames and inter-ethnic marriage. The analysis focuses on who gains and who loses (wage penalties for perceived Sephardic ethnicity) rather than on optimal resource allocation or aggregate welfare.

Category 5: Both efficiency and equity. **Autor, David H., and David Scarborough.** (2008). *“Does Job Testing Harm Minority Workers? Evidence from Retail Establishments.”*

The paper is framed as an equality–efficiency trade-off: it evaluates the productivity and retention gains from formal screening (an efficiency margin) alongside consequences for minority hiring (an equity margin), and interprets the findings through both lenses.

LLM explanation (gpt-5-mini-2025-08-07): The paper explicitly studies an “equality-efficiency trade-off”: it estimates the productivity effects of job testing (aggregate efficiency gains measured by tenure and turnover) and its distributional consequences for minority hiring (equity). The introduction and conclusion frame and interpret results in both lenses, finding productivity improvements without disparate impacts on minority hiring.

D.2 Economic Report of the President transmittal-letter scoring

Following Appendix C.2, each ERP transmittal letter is assigned two separate rubric-based scores on a 1–3 scale: an *efficiency score* and an *equity score*. The examples below illustrate how the scoring anchors map onto the substantive content and organizing logic of the President’s message. In each case, we briefly summarize the letter’s principal themes and then justify the assigned scores in the terms of the rubric.

1985: efficiency score = 3; equity score = 1. The 1985 transmittal letter is structured around a self-conscious defense of a supply-side and market-oriented policy program. It presents the administration’s economic agenda as a four-part strategy: restraining Federal spending growth, reducing personal and business tax rates, providing regulatory relief, and

sustaining a monetary policy oriented toward price stability and financial market health. The narrative emphasizes improved macroeconomic performance (lower inflation and interest rates, rising investment, and stronger productivity growth) and repeatedly frames policy as the removal of government-created impediments to private-sector dynamism, innovation, and long-run growth. It also highlights the fiscal deficit primarily as a spending problem, arguing that deficit reduction should proceed through expenditure restraint rather than tax increases, and it returns to tax reform as a vehicle for strengthening incentives for saving, investment, and productive activity.

Rationale for the scores: This letter earns an *efficiency score of 3* because efficiency-related themes (productivity, investment incentives, allocative improvements through deregulation, and an explicit privileging of policies aimed at expanding long-run productive capacity) form the central organizing logic of the message and are revisited across multiple policy domains (fiscal policy, taxation, regulation, and monetary policy). By contrast, it earns an *equity score of 1* because distributional and fairness-oriented claims are largely peripheral: the letter contains only limited, non-organizing references to unemployment, poverty, or protecting the safety net, without sustained attention to inequality, disadvantaged groups, or a set of explicitly redistributive or inclusion-focused policy commitments.

2017: efficiency score = 1; equity score = 3. The 2017 transmittal letter is organized around the proposition that macroeconomic recovery is insufficient unless its gains are broadly shared. It foregrounds labor-market outcomes for working families (job growth, rising wages, and reductions in poverty) and frames the administration’s legacy in terms of

expanding economic security and access, especially through health insurance coverage, education affordability, and stronger financial regulation. The letter repeatedly invokes the idea of “a fair chance” and emphasizes inclusion: it stresses that prosperity should not accrue only “at the top,” and it highlights policy accomplishments and priorities that are explicitly distributional in orientation (e.g., tax changes described as increasing the income share of the bottom 99 percent, expanded health coverage, and measures intended to strengthen worker bargaining power and job quality). While the letter also gestures toward long-run growth through investment in infrastructure and R&D, competition, innovation, and engagement with global trade, these points are presented as supportive complements rather than as the principal justificatory frame.

Rationale for the scores: This letter receives an *equity score of 3* because equity-related considerations are not merely present but central: the text repeatedly centers who benefits from growth, devotes substantial space to distributional outcomes (wages, poverty, inequality), and motivates policy priorities through the lens of inclusion and economic security for working- and middle-class families. Conversely, it receives an *efficiency score of 1* because, although the letter contains some genuine efficiency-adjacent content (productivity, competition, and investment), those themes do not structure the message and appear primarily as secondary supporting arguments rather than as a dominant emphasis on incentives, allocative efficiency, or long-run productive capacity.

1969: efficiency score = 3; equity score = 3. The 1969 transmittal letter is organized around a demanding two-objective program: sustaining high employment and growth while re-establishing price stability, and doing so in a manner that extends opportunity and reduces

poverty. On the efficiency side, the letter treats fiscal and monetary discipline as the “first line of defense” against inflation and frames the central macroeconomic task as keeping aggregate demand within the economy’s productive capacity. It then develops a distinct and sustained “Productivity and Efficiency” agenda, emphasizing improvements in the productive efficiency of industries and workers, the identification and removal of impediments to efficiency (with concrete sectoral examples), and a continuing commitment to competition, modernization, and policies that raise potential output.

Rationale for the scores: This combination warrants a *3 on efficiency* because efficiency is not a passing motif but a central organizing frame: the letter repeatedly returns to the imperative of noninflationary prosperity, explicitly links stabilization policy to the economy’s productive capacity, and devotes substantive discussion to productivity, industrial efficiency, and the policy-and-institutional reforms needed to lower costs and improve resource use. It likewise warrants a *3 on equity* because distributional objectives are elevated to a coequal national priority rather than appended as reassurance. The letter states, at the outset, that the nation must “share more equitably the fruits of prosperity” and it then spends extensive space detailing the social commitments meant to operationalize that aim, including worker protection, social insurance, anti-poverty and manpower programs, and policies explicitly motivated by opportunity for disadvantaged groups. In short, both lenses are repeatedly invoked, developed with concrete policy content, and used to structure the President’s account of the economic agenda, meeting the rubric’s “strong emphasis” standard for each dimension.

D.3 Policy-evaluation classification

Following Appendix C.3, each article is classified as either a *policy evaluation* (category = 1) or *not a policy evaluation* (category = 0). The examples below are designed to make the distinction concrete. For each category, we provide a representative paper, a brief justification that applies the definition in Appendix C.3, and then the corresponding LLM rationale recorded at the time of coding.

Category 1: Policy evaluation. Malamud, Ofer, and Cristian Pop-Eleches.

(2011). “*Home Computer Use and the Development of Human Capital.*”

This paper is organized around estimating the causal impact of a specific, well-defined public intervention (Romania’s Euro 200 voucher program) on a portfolio of child and household outcomes. The identification strategy is explicitly tied to program rules (the income cutoff) and the empirical exercise is, in essence, an estimate of what the voucher program *did*, relative to a credible counterfactual, which places it squarely within the policy-evaluation definition.

LLM explanation (gpt-5-mini-2025-08-07): The paper evaluates Romania’s Euro 200 government voucher program (a concrete public intervention) and uses a regression discontinuity design exploiting the income cutoff to estimate the causal impact of winning a voucher on computer ownership, use, academic achievement, cognitive and non-cognitive outcomes. This is a direct policy evaluation of the program’s effects.

Category 0: Not a policy evaluation. Dahlberg, Matz, Karin Edmark, and

Heléne Lundqvist. (2012). “*Ethnic Diversity and Preferences for Redistribution.*”

Although the paper leverages a refugee placement policy, the policy is not the estimand: it

functions primarily as a source of plausibly exogenous variation in local immigrant shares. The central causal object is the effect of ethnic diversity (immigrant presence) on natives' redistribution preferences, not the effect of the placement program itself; under Appendix C.3, this is therefore classified as *not* a policy evaluation.

LLM explanation (gpt-5-mini-2025-08-07): The study exploits a refugee placement program as an instrument to identify the causal effect of increased immigrant shares on natives' preferences for redistribution. Because the placement policy is used mainly as a source of exogenous variation (an instrument) and the object of interest is the effect of immigrant share rather than the policy itself, this does not meet the definition of a policy evaluation.