The Legibility Premium: Public Data Visibility and the Allocation of Competitive Federal Grants

Siqi Wei*
California State University, Northridge

Xiaoyang Zhu†
Wichita State University

May 2026

Abstract

We introduce the Public Data Visibility (PDV) index, a county-year measure of administrative legibility constructed from eight domains of publicly available federal data. A signal-extraction model predicts that federal agencies screen applicants by within-year visibility rank under bounded review capacity, so that the grant effect operates at the extensive margin of entering the considered set rather than through continuous precision improvements. Using a panel of 3,144 U.S. counties over fiscal years 2012–2022 and a hand-curated universe of 34 competitive federal grant programs, we find that a 10-percentile improvement in PDV rank is associated with a 2.2% increase in real per-capita competitive grant receipts. The effect concentrates in the Coverage sub-index and in domains aligned with funded programs (Schools, Environment, Health), and is entirely absent in formula grants, direct payments, and direct loans. Instrumental variables estimates confirm a positive causal effect. The findings provide quantitative evidence for Scott's (1998) legibility hypothesis in the contemporary U.S. federal allocation context, and they suggest that investments in county-level data infrastructure has meaningful allocative consequences.

JEL Codes: H77, H50, R51, D73, O10.
Keywords: public data visibility; legibility; federal grants; allocation.

*Email: siqi.wei@csun.edu.
†Email: xiaoyang.zhu@wichita.edu.

1 Introduction

The geography of federal resource allocation in the United States is strikingly uneven. Per capita federal discretionary grant receipts vary by more than two orders of magnitude across U.S. counties, and existing explanations of this variation appeal to a familiar set of factors: demographic targeting, political connections, statutory formulas, and administrative capacity. These factors plausibly account for much of the observed inequality. But they share a common assumption — that federal allocators are choosing among counties they can equally well see. This paper proposes that this assumption fails in an empirically meaningful way, and that the failure has substantial consequences for the distribution of competitive federal grants.

The unifying observation is that federal grant allocation is a decision problem under information constraints. A federal program officer deciding among applications for, say, an Economic Development Administration Public Works grant or an NIH research center grant cannot directly observe the underlying social or scientific value of work proposed in each applicant county. The officer relies on documentary evidence supplied by applicants and on indicators drawn from public data sources. A county whose conditions are richly documented in machine-readable federal datasets supplies a more informative signal than a county whose conditions appear in few datasets or in stale, low-resolution, or hard-to-parse forms. Whether deliberately or mechanically, the allocator weights the visible counties more heavily.

This intuition is not new in social theory. Scott (1998) identified "legibility" as a foundational logic of modern statecraft: states see what their measurement apparatus has been built to see, and this seeing shapes what states can do. What is new here is the attempt to quantify legibility at the U.S. county level, and to test whether it predicts the kinds of allocative outcomes Scott's framework implies.

The paper makes four contributions. The first contribution is conceptual and methodological. We introduce the Public Data Visibility (PDV) index, a county-year measure of administrative legibility that aggregates evidence from publicly available federal datasets across eight substantive domains: health, environment, broadband, housing, transportation, schools, local finance, and business/labor. The PDV scoring rubric is rule-based, programmatic, and reproducible, and the resulting panel covers 3,144 U.S. counties over fiscal years 2012–2022. The index is decomposable into three sub-dimensions (Coverage, Resolution, Usability) that capture distinct aspects of legibility.

The second contribution is a signal-extraction model that derives testable predictions about the relationship between PDV and federal grant allocation. In the model, a federal

officer maximizing the social value of a fixed competitive budget cannot directly observe true county need; the officer instead observes a noisy signal whose precision rises in the county's data visibility. Bayesian updating yields a posterior expected need that the agency would use as the basis for allocation if it could process every county. Bounded review capacity forces the agency to process only the top fraction of counties by visibility, where "top" is defined by within-year percentile rank rather than absolute visibility level. The model yields four predictions: the average effect of visibility on grants is strictly positive but operates entirely through the extensive margin of admission to the agency's processing set; the effect is detectable only for visibility's extensive-margin component and not for its intensive-margin sub-dimensions; only within-year visibility rank, not absolute level, determines allocation; and the effect is absent in grant categories where federal officer signal extraction does not occur (formula grants, direct entitlement payments, direct loans). A corollary of the threshold structure predicts further that the marginal effect of visibility is concentrated among mid-sized counties whose baseline visibility lies near the rank-based threshold.

The third contribution is empirical. Using a hand-curated panel of 34 truly competitive federal grant programs spanning eleven federal agencies, we show that a 10-percentile improvement in a county's within-year PDV rank is associated with approximately a 2.2% increase in real per-capita competitive grant obligations. The effect is concentrated in the Coverage sub-index, consistent with the model's screening-threshold interpretation: what matters is whether a county is visible at all, not how granular or technically polished the underlying data are. The effect is concentrated in the domains aligned with the funded programs (Schools, Environment, Health). Crucially, the effect is entirely absent in placebo outcomes: formula grants, direct entitlement payments, and direct loans show no relationship with PDV, ruling out generic state capacity or political-clout confounding as the operative mechanism. The underlying continuous PDV index in its raw form produces a small negative and statistically insignificant coefficient on the same outcome, confirming that the operative relationship is rank-based rather than level-based. Instrumental variables estimates exploiting the staggered adoption of state open-data laws and a Bartik shift-share predictor confirm a positive causal effect with first-stage F-statistics above 15 in all specifications.

The fourth contribution is policy-relevant. The decomposition of the visibility effect onto the extensive margin — the binary fact of being measurable — identifies a specific policy lever. Federal and state investments in data infrastructure that bring previously invisible counties across the measurement threshold should have larger allocative consequences than investments that refine already-visible data. Aggregated to the population scale, the 11% proportional change associated with a movement across the interquartile range of the within-year PDV distribution implies approximately $6 per capita in additional competitive grants

when evaluated at the panel-wide unconditional mean of $54 per capita, and approximately $13 per capita when evaluated at the conditional mean of $122 per capita among the 44% of county-years with positive competitive grant receipts. For a county of fifty thousand residents this amounts to between $300,000 and $650,000 annually in additional federal investment, depending on whether the county begins as an occasional or as an established recipient.

The remainder of the paper is organized as follows. Section 2 situates the paper in three literatures. Section 3 develops the theoretical model and derives the testable predictions. Section 4 describes the construction of the PDV index and the competitive grants panel. Section 5 sets out the econometric strategy. Section 6 reports the empirical results and discusses their interpretation. Section 7 concludes.

2 Related Literature

The paper draws on and contributes to three distinct literatures. The first is the literature on state capacity and legibility. The foundational treatment of legibility as a logic of statecraft is Scott (1998), who argued that modern states create the measurement infrastructure that lets them see populations, and that this seeing both enables targeted intervention and reshapes the populations seen. Subsequent work in development economics, public finance, and political economy has formalized state capacity and documented its consequences for economic outcomes (Besley and Persson (2009); Acemoglu et al. (2015, 2016)). This literature has emphasized measurable capacity — fiscal penetration, bureaucratic effectiveness, statutory enforcement — but has paid less attention to the informational substrate that capacity operates through. The PDV measure introduced here provides a direct empirical analogue to Scott's notion of legibility in the contemporary U.S. federal context.

The second is the empirical literature on the geography of federal resource allocation. Suárez Serrato and Wingender (2016) estimated local fiscal multipliers using variation in federal spending across U.S. counties. Chodorow-Reich (2019) surveyed the methodological challenges in geographic identification of fiscal effects. Adelino et al. (2017) studied municipal bond rating recalibration and the resulting reallocation of federal benefits. Cellini et al. (2010) estimated the effects of school-bond approval on house prices. This literature typically takes the allocation of federal funds as exogenous (or exogenizes it through formula discontinuities) and estimates downstream effects. The present paper inverts the direction: it asks why federal allocations land where they do, and provides evidence that the documentary infrastructure of each potential recipient is a meaningful factor.

The third is the literature on measurement inequality and the representation of places in public data. Chetty et al. (2014) and related work document substantial geographic het-

erogeneity in opportunity-relevant outcomes and develop new measures from administrative data. Census research on the differential undercount establishes that some places are systematically less well-measured than others (Hogan (2003)). The PDV index contributes to this literature a direct measure of administrative visibility — not the underlying conditions, but the documentation of conditions — and shows that this measure has allocative consequences.

A related strand of work in the political science and public administration literatures has studied open-data initiatives and the "open government" movement (Ruijer et al. (2020); Harrison et al. (2012)). This work has largely been institutional and descriptive, documenting state and local open-data policies and their adoption patterns. The contribution of the present paper is to provide a quantitative measure of the resulting data environment and a credible estimate of its causal effects on federal allocation.

3 A Model of Visibility-Mediated Federal Grant Allocation

This section develops a signal-extraction model of federal grant allocation in which the awarding agency observes a noisy signal of each county's underlying need, with the precision of the signal depending on the county's data visibility. Under bounded attention, the agency can process signals only from a top fraction of counties in the within-year visibility distribution. The model yields four predictions that the empirical analysis subsequently tests, all of which fall out of the equilibrium allocation rule rather than being imposed as primitives.

3.1 Setup

A federal agency has a fixed competitive budget $B > 0$ to allocate across a continuum of counties indexed by $i \in [0, 1]$. Each county is characterized by two primitives:

- **True need** θ_i, drawn from a distribution F_θ on $[0, \overline{\theta}]$ with mean μ_θ and precision τ_θ (the inverse of the prior variance). The parameter θ_i summarizes the social value of allocating grant dollars to county i — equivalently, the marginal social return on federal investment in that county.

- **Visibility** $v_i \in [0, 1]$, drawn from a distribution F_v that is independent of θ_i. The parameter v_i captures the extent to which county i's conditions are documented in publicly available federal datasets. A county with $v_i = 1$ is fully documented; a county with $v_i = 0$ is invisible to the federal agency.

True need θ_i is private information to the county; the agency must infer it from publicly observable evidence whose informativeness depends on v_i.

3.2 Information Structure: Signal Extraction with Visibility-Dependent Precision

For each county that the agency processes, it observes a noisy signal of true need,

$$s_i = \theta_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\big(0,\, \sigma_0^2 \,/\, v_i\big)\,, \tag{1}$$

where $\sigma_0^2 > 0$ is a baseline noise scale. The signal *precision* is $\tau(v_i) = v_i/\sigma_0^2$, increasing linearly in visibility. As $v_i \to 0$ the noise variance diverges and the signal carries no information; as $v_i \to 1$ the signal becomes maximally informative. The multiplicative form $\sigma^2(v_i) = \sigma_0^2/v_i$ has a natural microeconomic interpretation: a fraction v_i of the true conditions in county i is documented in public data, and the remaining $(1 - v_i)$ generates measurement noise of fixed magnitude. The exact functional form is not essential to the qualitative predictions of the model.

Given the prior $\theta_i \sim \mathcal{N}(\mu_\theta, \tau_\theta^{-1})$ and the signal in equation (1), the agency's posterior expectation by Bayes' rule is

$$m_i \equiv \mathbb{E}[\theta_i \,|\, s_i, v_i] = \frac{\tau_\theta\, \mu_\theta + \tau(v_i)\, s_i}{\tau_\theta + \tau(v_i)} = \mu_\theta + \frac{\tau(v_i)}{\tau_\theta + \tau(v_i)}(s_i - \mu_\theta). \tag{2}$$

The posterior weight on the signal, $\tau(v_i)/[\tau_\theta + \tau(v_i)]$, is monotonically increasing in v_i. Counties with high visibility have posteriors that track their signals closely; counties with low visibility have posteriors close to the prior mean.

3.3 Bounded Attention and the Rank-Based Threshold

The agency cannot process signals from all counties. Processing a signal requires costly review, and the agency's review capacity is bounded: it can rigorously evaluate only a top fraction $\alpha \in (0, 1)$ of applicants in each fiscal cycle. The optimal selection under linear utility is to rank counties by the precision of their available signal — equivalently, by their visibility — and process the top-α fraction.[1]

The implied screening threshold is the within-year $(1 - \alpha)$-th percentile of the visibility

[1] A formal derivation of this selection rule follows from standard bounded-attention arguments (Sims, 2003; Caplin and Dean, 2015): the marginal value of an additional signal is proportional to its precision, and a fixed review-capacity constraint implies that the agency processes signals in descending order of precision.

distribution:

$$\underline{v}_t \;=\; F_{v,t}^{-1}(1-\alpha). \tag{3}$$

Counties with $v_{i,t} \geq \underline{v}_t$ enter the agency's *visible set* $\mathcal{V}_t$ and have their signals processed; counties with $v_{i,t} < \underline{v}_t$ are excluded.

The fact that $\underline{v}_t$ is defined by a within-year rank, not by an absolute level on the visibility scale, is the model's central methodological implication. The threshold $\underline{v}_t$ moves mechanically as the distribution $F_{v,t}$ shifts over time — for instance, as national data infrastructure expands and the floor of "minimally documented" rises. A county whose raw visibility improves at the national rate experiences no change in its position relative to the threshold; only counties whose visibility improves *faster* than the national distribution gain entry into $\mathcal{V}_t$.

This rank-based threshold structure is also consistent with the institutional features of modern federal allocation tools: the Climate and Economic Justice Screening Tool, USDA's Persistent Poverty Counties designation, the IRA Energy Communities provisions, and HUD's distress indices all use within-year percentile thresholds to determine eligibility for consideration, mirroring the threshold $\underline{v}_t$ in equation (3).

3.4 Allocation Rule

For counties in the visible set, the agency allocates the budget proportionally to posterior expected need:

$$g_i \;=\; \begin{cases} B \cdot \dfrac{m_i}{D_t} & \text{if } i \in \mathcal{V}_t, \\ 0 & \text{if } i \notin \mathcal{V}_t, \end{cases} \tag{4}$$

where $D_t \equiv \int_{\mathcal{V}_t} m_j \, dj$ is the integrated posterior expected need across the visible set.[2]

3.5 Comparative Statics and Testable Predictions

The four propositions that follow are now derived from the allocation rule in equation (4).

Proposition 1 (Visibility raises competitive grant capture)**.** *The expected partial effect of a county's visibility on its grant allocation, averaged across the joint distribution of* (θ_i, v_i)*, is strictly positive:*

$$\mathbb{E}\left[\frac{\partial g_i}{\partial v_i}\right] \;>\; 0. \tag{5}$$

[2]Proportional allocation is the welfare-maximizing rule when the agency's utility is linear in $\theta_i g_i$ and it must commit to a smooth allocation across applicants under a budget constraint. The qualitative predictions also obtain under a top-K rule in which the agency funds the highest-posterior applicants up to budget exhaustion.

The positive average effect operates through the extensive margin of admission to the visible set.

Proof. Compute the expected allocation as a function of visibility. For $v_i < \underline{v}_t$, $g_i = 0$ and so $\mathbb{E}[g_i \mid v_i] = 0$. For $v_i \geq \underline{v}_t$, take the expectation of equation (4) over the joint distribution of the signal s_i and other counties' realizations. Two facts together pin down the result. First, by independence of θ_i and v_i and the zero-mean noise in equation (1),

$$\mathbb{E}[m_i \mid v_i] \;=\; \mu_\theta + w(v_i) \cdot \mathbb{E}[s_i - \mu_\theta \mid v_i] \;=\; \mu_\theta,$$

where $w(v_i) = \tau(v_i)/(\tau_\theta + \tau(v_i))$. The posterior *variance* falls with v_i, but the posterior *mean* equals the prior mean once we marginalize over the unobserved θ_i. Second, by the law of large numbers across the continuum of counties in $\mathcal{V}_t$, $D_t = \int_{\mathcal{V}_t} m_j \, dj \to \alpha\mu_\theta$ almost surely. Combining,

$$\mathbb{E}[g_i \mid v_i] \;=\; \frac{B}{\alpha} \cdot \mathbf{1}\{v_i \geq \underline{v}_t\}.$$

The expected allocation is therefore a step function of visibility: zero below the threshold and B/α above. The average partial effect across the visibility distribution is

$$\mathbb{E}_v\left[\frac{\partial \mathbb{E}[g_i \mid v_i = v]}{\partial v}\right] \;=\; \frac{B}{\alpha} \cdot f_v(\underline{v}_t) \;>\; 0,$$

strictly positive because the density at the threshold is strictly positive. The aggregate expected effect of visibility on grants operates entirely through the extensive margin of admission to the visible set: signal-precision improvements within the visible set reallocate dollars across counties (sharpening high-need posteriors upward and low-need posteriors downward) but cannot raise the average expected allocation, which is pinned at B/α by the budget constraint. □

The proof reveals that Proposition 1 is not a mechanical consequence of the model's functional form but rather a substantive prediction about which margin of visibility carries the empirical effect. Within the visible set, signal sharpening reallocates dollars from below-average-need to above-average-need counties but does not raise the aggregate flow to any particular subset. The aggregate positive effect of visibility on grants comes *entirely from threshold crossing* — counties moving from exclusion to inclusion in the visible set.

Proposition 2 (Extensive margin: the visibility threshold)**.** *A county i in year t receives a strictly positive grant allocation if and only if its visibility crosses the within-year threshold:*

$$\Pr(g_{i,t} > 0 \,|\, v_{i,t}) \;=\; \mathbf{1}\{v_{i,t} \geq \underline{v}_t\}. \tag{6}$$

The unconditional probability of receiving any grant in year t is exactly α, the agency's attention budget.

Proof. Immediate from equation (4) and the definition of the visible set $\mathcal{V}_t$ in equation (3). □

Proposition 2 predicts that the extensive margin of visibility — whether the county is in the agency's processing set at all — is a precondition for receiving allocations. The extensive-margin variation corresponds to the Coverage sub-index of the PDV measure constructed in Section 4; the intensive-margin variation (signal-precision improvements *within* the visible set) corresponds to the Resolution and Usability sub-indices.

Proposition 3 (Rank matters, not level)**.** *The expected grant allocation $\mathbb{E}[g_{i,t} \mid v_{i,t}]$ depends on visibility only through its within-year percentile rank $F_{v,t}(v_{i,t})$. Two counties with identical within-year rank but different absolute visibility have identical expected allocations.*

Proof. The proof of Proposition 1 established that $\mathbb{E}[g_{i,t} \mid v_{i,t}] = (B/\alpha) \cdot \mathbf{1}\{v_{i,t} \geq \underline{v}_t\}$. Substituting the rank-based threshold $\underline{v}_t = F_{v,t}^{-1}(1-\alpha)$ from equation (3),

$$\mathbb{E}[g_{i,t} \mid v_{i,t}] \;=\; \frac{B}{\alpha} \cdot \mathbf{1}\{F_{v,t}(v_{i,t}) \geq 1-\alpha\},$$

which depends on $v_{i,t}$ only through its within-year percentile rank $F_{v,t}(v_{i,t})$. Two counties with identical rank therefore have identical expected allocations regardless of their absolute visibility levels. □

Proposition 3 is a substantive prediction of the bounded-attention specification: because the agency's processing budget is fixed in rank terms, the absolute level of visibility is informationally redundant. A county whose raw PDV improves at the national rate but maintains its rank gains no allocative advantage. The empirical analysis tests this prediction by reporting the headline regression separately with raw PDV, z-scored PDV, and within-year percentile rank as the regressor.

Proposition 4 (Mechanism specificity)**.** *Let g_i^{comp} denote allocation under the competitive program described above. Let $g_i^{form} = h(\theta_i, x_i)$ denote allocation under a formula-grant program in which the allocation function h depends on statutory inputs x_i (population, poverty rate, road miles, etc.) that are independent of v_i conditional on θ_i. Let $g_i^{dir} = c \cdot N_i$ denote direct entitlement payments to individuals based on eligible population N_i. Then*

$$\mathbb{E}\left[\frac{\partial g_i^{comp}}{\partial v_i}\right] > 0, \quad \frac{\partial g_i^{form}}{\partial v_i} = 0, \quad \frac{\partial g_i^{dir}}{\partial v_i} = 0. \tag{7}$$

Proof. The competitive case is Proposition 1. The formula and direct-payment cases follow immediately: $h(\theta_i, x_i)$ and $c \cdot N_i$ do not depend on v_i because the corresponding allocation mechanisms make no use of agency signal extraction. □

Proposition 4 predicts that the visibility effect is specific to programs in which the agency performs signal extraction. Formula grants are allocated by exogenous statutory rules that do not depend on documented need. Direct entitlement payments flow to individuals based on eligibility, not county-level documentation. The placebo battery in Section 6 tests this prediction by running the headline regression on formula grants, direct payments, and direct loans alongside competitive grants.

3.6 An Endogenous Corollary: Concentration at the Threshold

The step-function form of $\mathbb{E}[g_i \mid v_i]$ established in the proof of Proposition 1 has a sharp implication for cross-county heterogeneity. The effect of a marginal increase in visibility on a county's expected allocation depends entirely on whether that increase pushes the county across the threshold $\underline{v}_t$. For counties well above the threshold, additional visibility has no expected effect; the county is already in the visible set and the budget-balanced posterior mean is invariant to further precision gains. For counties well below the threshold, additional visibility also has no expected effect; the county remains excluded. Only counties whose baseline visibility lies near the threshold see an expected allocation change from a small visibility improvement.

Corollary 1 (Concentration at the threshold)**.** *For any $\delta > 0$,*

$$\mathbb{E}[g_i \mid v_i + \delta] - \mathbb{E}[g_i \mid v_i] \;=\; \begin{cases} B/\alpha & \text{if } v_i < \underline{v}_t \leq v_i + \delta, \\ 0 & \text{otherwise.} \end{cases}$$

The change in expected allocation from a small visibility improvement is therefore positive only for counties whose baseline visibility lies in the interval $[\underline{v}_t - \delta, \underline{v}_t)$, and zero for all other counties.

In the empirical specification, Corollary 1 predicts that the visibility–grants relationship should be detectable only in counties whose baseline visibility lies near the threshold $\underline{v}_t$. Because larger counties tend to have higher baseline visibility for mechanical reasons (more federal data collection occurs in more populous places) and smaller counties have lower baseline visibility, the population layer near the threshold falls in the middle of the county size distribution. This generates the heterogeneity finding documented in the empirical section:

the effect is concentrated in the second population quartile, with weaker and statistically indistinguishable effects in the smallest and largest quartiles. The corollary turns this from a post-hoc interpretation into a formal prediction of the model.

3.7 Mapping to Empirical Specifications

Translating the model into estimable form requires linking the unobserved theoretical v_i to the measured PDV index. The empirical analysis treats the within-year percentile rank of a county's PDV composite, $\text{PDV}_{i,t}^{\text{pct}}$, as the empirical analog of the rank-relative position implied by the model's threshold structure. True need θ_i is partially captured by observable demographic and economic controls and absorbed via county fixed effects in the panel specifications. Year fixed effects absorb common shifts in the visibility distribution. State-by-year fixed effects absorb state-level shocks to the competitive grant allocation environment. The headline regression estimates the empirical analog of $\mathbb{E}[\partial g_i/\partial v_i]$ from Proposition 1, the sub-index decomposition tests the Coverage-versus-intensive-margin prediction of Proposition 2, the raw-versus-rank specifications test Proposition 3, the placebo battery across grant categories tests Proposition 4, and the population quartile heterogeneity tests Corollary 1.

4 Data and Variable Construction

The empirical analysis uses two principal data assets, both constructed for this paper and described in detail in the technical appendices. This section summarizes the construction and reports descriptive statistics.

4.1 The Public Data Visibility Index

The Public Data Visibility (PDV) index is a county-year measure of administrative legibility constructed from publicly available federal datasets. It is the empirical analog of the visibility primitive v_i in the model of Section 3.

Conceptual structure. PDV aggregates evidence across eight substantive domains chosen to span the major policy areas in which federal grants flow: health, environment, broadband, housing, transportation, schools, local finance, and business and labor. Within each domain, a county receives a score on a five-point ordinal rubric from 0 to 4 based on whether public federal data documenting the county's conditions in that domain exist, at what spatial resolution, in what machine-readable form, and with what currency. A score of 0 indicates

that no public data document the county in the domain; a score of 4 indicates that sub-county data are available, machine-readable, published via a documented API or FTP feed, updated within two years, and linkable via stable FIPS or GEOID identifiers.

Composite and sub-indices. The composite PDV index for county i in year t is the simple average of the eight domain scores:

$$\mathrm{PDV}_{i,t}^{\mathrm{raw}} = \frac{1}{8} \sum_{d=1}^{8} S_{i,d,t}, \tag{8}$$

where $S_{i,d,t} \in \{0, 1, 2, 3, 4\}$ is the rubric score for county i in domain d in year t. The composite is decomposed into three sub-indices that capture distinct aspects of legibility: **Coverage** (the share of domains in which the county has any measurable data, capturing the extensive margin of visibility), **Resolution** (the average spatial granularity of available data), and **Usability** (the average technical accessibility of the data).

Panel coverage. The PDV panel covers all 3,144 counties and county-equivalents in the 50 states and the District of Columbia, over fiscal years 2012–2022, for a total of 34,574 county-year observations. The underlying data sources span all major federal statistical and regulatory agencies; the complete source registry, scoring rubric, and domain-level construction are documented in Appendix A.

National evolution of the index. The raw PDV composite rises monotonically over the panel period from a national mean of 2.79 in 2012 to 3.75 in 2022, an increase of 34.6% in eleven years. The trajectory is not smooth: three discrete steps account for most of the cumulative growth. Between 2015 and 2016 the mean jumps from 3.01 to 3.52, reflecting the launch of CDC PLACES (initially the 500 Cities Project), which made tract-level chronic disease data available for the 339 most populous counties and county-level estimates available for the remainder. Between 2019 and 2020 the mean jumps from 3.65 to 3.75, reflecting the nationwide expansion of CDC PLACES to all U.S. counties. Smaller increments in 2014, 2017, and 2022 reflect, respectively, the launch of EPA EJScreen and FCC Form 477 broadband reporting, EPA's expanded TRI release, and the launch of the FCC Broadband Data Collection with its sub-block-level API. Figure 1 plots the national mean alongside the within-year inter-quartile and P10–P90 ranges of county-level PDV. The shaded bands narrow substantially over the panel, particularly after 2016, indicating that the cross-county dispersion in visibility has fallen even as the national mean has risen.

Sub-index and domain composition. The composite movement masks substantial differences across sub-indices and domains. Table 1 reports the year-by-year means of the three sub-indices. Coverage rises sharply between 2015 and 2016 (from 0.87 to 0.99) and effectively saturates thereafter, reflecting that by 2016 nearly every county is measurable in nearly every domain. Resolution rises more gradually (from 2.07 to 3.05), reflecting incremental improvements in spatial granularity across multiple domains. Usability rises in two discrete steps (in 2016 and 2017) corresponding to the introduction of machine-readable APIs for the largest federal datasets. The Coverage and Usability sub-indices have markedly lower cross-county dispersion than the Resolution sub-index because whether a county is measurable at all and whether the data are technically accessible largely move at the national level, while spatial granularity varies more across counties at a point in time. At the domain level, Health and Transportation exhibit the largest cross-county heterogeneity (driven by the geographic incidence of monitoring infrastructure), while Broadband and Business/Labor exhibit minimal cross-county variation in any given year because the underlying data are universally available or universally limited.

Within-year percentile transformation. The within-year percentile transformation, denoted $\text{PDV}_{i,t}^{\text{pct}}$, has a mean of approximately 50 and a standard deviation of approximately 28 by construction. The transformation strips out the national-trend component of PDV that year fixed effects already absorb in the panel specifications and isolates each county's relative position in the within-year distribution. The motivation for using the percentile transformation rather than the raw composite as the regressor of interest is Proposition 3, which establishes that the relevant determinant of competitive grant allocation is the county's within-year rank rather than its absolute level on the visibility scale.

Figure 2 presents the geographic distribution of the PDV composite score across U.S. counties. Higher-score areas are primarily concentrated along the West Coast, while counties in the Midwest tend to exhibit relatively lower scores. Figure 3 re-expresses each county's PDV score as a within-year percentile rank relative to all county-year observations in the same calendar year, and then averages these percentile ranks across years. As shown, the spatial distribution pattern is broadly similar to that of the composite score. Figure 4 plots the median county-level PDV score for each state, ordered from lowest to highest along the horizontal axis. Among the 50 states, Alaska, Iowa, South Dakota, North Dakota, and Nebraska exhibit the five lowest PDV scores, whereas Arizona, Maryland, New Jersey, California, and the District of Columbia display the highest scores.

4.2 The Competitive Grants Panel

The outcome variable is the panel of competitive federal grant obligations to U.S. counties, restricted to a hand-curated set of 34 federal grant programs awarded through peer review or discretionary competitive selection. This panel is the empirical analog of the competitive allocation g_i^{comp} in the model of Section 3.

Program universe. The 34 programs span eleven federal agencies and three broad areas: research grants administered through peer review (NSF directorates, NIH institutes), discretionary economic development and infrastructure grants (EDA Public Works, DOT BUILD/RAISE, EPA Brownfields, FEMA BRIC pre-disaster mitigation), and competitive health, education, and community programs (HRSA Health Center expansion, SAMHSA regional and national projects, HUD Choice Neighborhoods, ED IES education research, USDA Rural Development competitive grants, DOJ Second Chance Act). Programs were admitted to the universe on three criteria: federal officers exercise substantial discretion in recipient selection; the program receives more applications than it can fund and applicants compete on the merit of proposals; and the program disbursed at least $5 million nationally in a typical fiscal year between 2012 and 2022. The complete program list is reported in Appendix B together with the selection criteria and the data pull methodology.

Geographic attribution. Each award is attributed to the county where the funded activity occurs (place of performance), not the county where the legal recipient is headquartered (recipient location). This choice ensures that pass-through awards from state agencies are attributed to the counties where work actually occurs, consistent with the paper's research question about where federal money is geographically directed for use.

Variable construction. The headline outcome is the inverse hyperbolic sine of real per-capita competitive grant obligations:

$$y_{i,t} = \operatorname{arcsinh}\left(\frac{G_{i,t} \cdot (\text{CPI}_{2022}/\text{CPI}_t)}{N_{i,t}}\right), \tag{9}$$

where $G_{i,t}$ is the sum of nominal-dollar obligations from the 34 programs in fiscal year t, $N_{i,t}$ is the Census Population Estimates Program annual county population, and the CPI ratio converts to constant 2022 dollars. The inverse hyperbolic sine is selected because approximately 56% of county-year cells have $G_{i,t} = 0$ (the majority of small and rural counties win no competitive grants in a given year), and the logarithm would drop these observations (Bellemare and Wichman, 2020).

Panel coverage and descriptive moments. The competitive grants panel is balanced: 3,144 counties × 11 fiscal years = 34,574 observations. Approximately 44% of county-year cells record positive competitive grant obligations. Total pooled obligations over the panel period are approximately $281 billion. Mean per-capita real obligations are $54, with a median of zero and a maximum of approximately $19,269 in a single county-year cell.

Composition by agency and program. The pooled $281 billion is concentrated in a small number of federal funders. The National Institutes of Health contribute 49.0% of pooled obligations ($138.0 billion across ten institutes), the National Science Foundation contributes 26.9% ($75.8 billion across ten directorates), and a single HRSA program — the Health Center Expansion (CFDA 93.527) — contributes a further 13.5% ($38.0 billion). Together these three funders account for 89.4% of pooled dollar volume; the remaining eight federal agencies contribute the residual 10.6%. Substantively, the panel is dominated by research funding (76.0% of pooled dollars), with health, education, and community programs (18.9%) and economic development and infrastructure (5.1%) accounting for the balance. Figure 6 displays the annual dollar volume by agency, showing both the steady expansion of competitive grant volume over the panel period and the persistent dominance of biomedical research funding within the competitive universe. The full agency-by-agency and program-by-program breakdown is reported in Appendix B.

Geographic concentration of receipts. Competitive grant capture is highly concentrated geographically. The ten counties with the largest pooled obligations capture 27.2% of all dollar volume; the top fifty counties capture 59.5%; the top one hundred capture 75.1%. The Gini coefficient across all 3,144 counties (including 614 counties with zero pooled obligations) is 0.928, indicating an extreme degree of geographic concentration. At the state level, California alone receives 13.7% of pooled obligations; the three most research-intensive states (California, New York, Massachusetts) together capture 29.1%. This concentration is important for the empirical analysis: it implies that cross-sectional variation in per-capita grant capture is dominated by the right tail of the distribution, motivating the inverse hyperbolic sine transformation of the outcome that compresses the tail while preserving the zero–positive contrast. For more details, please see Table 7 in the Appendix B.

Persistence in grant receipt. Whether a county receives competitive grants in one fiscal year is strongly predictive of whether it receives them in the next. The conditional probability that a county receives any competitive grant in year t given that it received one in year $t-1$ is 0.84; the conditional probability of receipt given non-receipt in $t-1$ is 0.25.

Of the 3,144 counties in the panel, 18.3% received competitive grants in all eleven years and only 1.7% (54 counties) received nothing in any year; the remaining 80% received intermittently (Details see Table 9 in Appendix B). The strong persistence of receipt is consistent with the visibility-based screening process the paper formalizes: once a county enters the recipient pool, the documentary capacity generated by receiving and reporting on awards reinforces its subsequent visibility, while counties that have not entered face a structurally lower probability of crossing the visibility threshold in subsequent years.

4.3 Controls and Instruments

The empirical analysis uses a standard set of time-varying county controls drawn from publicly available federal sources: log population (Census PEP), unemployment rate (BLS LAUS), poverty rate (Census SAIPE), shares of population by race and ethnicity (Census PEP), share aged 65 or older (Census PEP), real median household income (Census SAIPE, CPI-deflated), Democratic presidential vote share (MIT MEDSL, carried forward between presidential election years), and the count of major and emergency disaster declarations (FEMA OpenFEMA). Two instruments are used in the causal identification section: an indicator for state-level open-data law adoption (coded from the National Conference of State Legislatures tracker and state executive orders) and an indicator for state Chief Data Officer position creation (coded from the Beeck Center State CDO Tracker). Construction details for all auxiliary variables are documented in the Appendix B.

5 Empirical Strategy

The empirical analysis tests the predictions of Section 3 through a coordinated set of panel specifications. The headline is a two-way fixed effects regression in which the within-year percentile rank of PDV is the regressor of interest. A long-difference design uses only the eleven-year panel endpoints as a robustness check on within-county variation. Two instrumental variables strategies address remaining endogeneity concerns. A placebo battery tests the model's prediction of mechanism-specificity to competitive grants. Sub-index and domain decompositions identify which aspects of visibility drive the relationship.

5.1 Baseline Specification

The headline regression follows from the linearized analog of the model's allocation rule:

$$y_{i,t} = \beta \cdot \mathrm{PDV}^{\mathrm{pct}}_{i,t} + \mathbf{x}'_{i,t}\boldsymbol{\gamma} + \mu_i + \nu_{s(i),t} + \varepsilon_{i,t}. \tag{10}$$

Here $y_{i,t}$ is the inverse hyperbolic sine of real per-capita competitive grant obligations in county i, fiscal year t; $\text{PDV}_{i,t}^{\text{pct}}$ is the within-year percentile rank of PDV; $\mathbf{x}_{i,t}$ is the vector of time-varying controls described in Section 4; μ_i is a county fixed effect; $\nu_{s(i),t}$ is a state-by-year fixed effect; and $\varepsilon_{i,t}$ is an idiosyncratic error term clustered at the county level. The county fixed effect absorbs all time-invariant unobservables, including baseline administrative capacity. The state-by-year fixed effect absorbs state-level shocks to allocation. The coefficient β is the empirical analog of $\partial g_i / \partial v_i$ in Proposition 1.

Interpretation of β. With $\text{PDV}_{i,t}^{\text{pct}}$ on a 0–100 scale and the inverse-hyperbolic-sine-transformed outcome, β is the approximate proportional change in real per-capita competitive grant obligations associated with a one-percentile-point increase in within-year PDV rank. Substantively meaningful effect sizes are reported as 10β (a 10-percentile-rank improvement) or 25β (an interquartile shift).

Choice of percentile rank rather than raw level. The within-year percentile transformation is motivated by Proposition 3: in a system where the federal allocation apparatus ranks counties against each other within each year, the relevant determinant of allocation is the county's position in the within-year distribution rather than its absolute level on the visibility scale. This methodological choice is testable: the empirical analysis in Section 6 reports the specification using raw PDV, z-scored PDV, and percentile rank in parallel.

5.2 Long-Difference Specification

As a within-county robustness check using only the panel endpoints, the analysis also estimates

$$\Delta y_i = \beta^{\text{LD}} \cdot \Delta \text{PDV}_i^{\text{pct}} + \Delta \mathbf{x}_i' \boldsymbol{\gamma} + \nu_{s(i)} + \varepsilon_i, \tag{11}$$

where Δ denotes the 2022-minus-2012 first difference and $\nu_{s(i)}$ is a state fixed effect. Standard errors are clustered at the state level. The long-difference specification uses different identifying variation than equation (10) (the eleven-year change rather than all year-on-year movements) and provides a robustness check on the headline. Figure 5 illustrates the geographic distribution of the ten-year change in PDV across U.S. counties. The largest improvements are primarily concentrated in the Midwest region.

5.3 Instrumental Variables

The TWFE specification controls for time-invariant county heterogeneity through county fixed effects, but PDV could remain endogenous to grant capture through time-varying

unobservables. Two identification concerns are particularly relevant. First, federal grants come with reporting requirements that themselves generate publicly visible data, raising the possibility of reverse causality. Second, time-varying unobserved state capacity could drive both PDV adoption and grant-writing competence.

State open-data law instrument. The analysis instruments $\text{PDV}_{i,t}^{\text{pct}}$ with the staggered adoption of state-level open-data laws and Chief Data Officer positions, both of which directly mandate or encourage the public release of administrative data and thereby shift county-level PDV. The first-stage regression takes the form

$$\text{PDV}_{i,t}^{\text{pct}} = \pi_1 Z_{s(i),t}^{\text{ODL}} + \pi_2 Z_{s(i),t}^{\text{CDO}} + \mathbf{x}_{i,t}'\boldsymbol{\delta} + \mu_i + \tau_t + u_{i,t}, \tag{12}$$

and the second stage replaces $\text{PDV}_{i,t}^{\text{pct}}$ in equation (10) with the predicted value. The exclusion restriction requires that state open-data laws affect competitive grant capture only through their effect on PDV, conditional on county and year fixed effects.

Bartik shift-share instrument. A second instrument exploits the differential exposure of counties to national PDV trends across domains. For each domain d, the national mean of the PDV score in year t exhibits substantial movement driven by federal data infrastructure investments. Counties with higher baseline exposure to growing domains see larger predicted PDV increases. The shift-share instrument is constructed as

$$Z_{i,t}^{\text{B}} = \sum_{d=1}^{8} \omega_{i,d,2012} \cdot \left(\overline{S}_{d,t} - \overline{S}_{d,2012}\right), \tag{13}$$

where $\omega_{i,d,2012}$ is county i's 2012 share of its composite PDV attributable to domain d, and $\overline{S}_{d,t}$ is the national mean of the domain d score in year t. Identification under the Goldsmith-Pinkham et al. (2020) framework rests on the exogeneity of the baseline shares, which is plausible because pre-2012 domain composition predates the federal data infrastructure expansions of the panel period.

5.4 Placebo Battery and Decompositions

The model's Proposition 4 predicts that the visibility effect is specific to competitive grants. To test this, the analysis estimates equation (10) across six different federal grant outcomes: competitive grants (the headline outcome), broad-classified discretionary grants (which include quasi-formula programs that pollute the bucket), formula grants, direct entitlement payments to individuals, direct loans, and total federal assistance. The model predicts a

positive coefficient only for the competitive category. To identify which aspects of visibility drive the headline relationship, the analysis additionally estimates two decompositions. The sub-index decomposition replaces the composite PDV percentile with the percentile ranks of the three sub-indices — Coverage, Resolution, and Usability — entered individually and jointly. The domain decomposition replaces the composite with the percentile ranks of each of the eight domain scores, again individually and jointly.

5.5 Heterogeneity and Sample Splits

The heterogeneity analysis estimates the headline specification separately on subsamples defined by three dimensions of county characteristics. Population quartiles are computed within each fiscal year using Census PEP annual county population, with cutoffs at the 25th, 50th, and 75th within-year percentiles — so a county may shift quartiles across years as its relative size changes. Era splits are defined by two policy dates: the Digital Accountability and Transparency Act's full compliance deadline of May 2017, with pre-DATA-Act years being 2012–2016 and post-DATA-Act years 2017–2022, and the onset of the COVID-19 pandemic, with pre-COVID years being 2012–2019 and post-COVID years 2020–2022. Racial-composition splits use the Census PEP shares of Black (non-Hispanic) and Hispanic populations and the AIAN-alone share. A county is classified as high-Black if its Black share exceeds 15%, high-Hispanic if its Hispanic share exceeds 15%, and high-AIAN if its AIAN share exceeds 5%. The corresponding low-share subsamples are the complements — counties below the cutoffs in each demographic dimension.

Two additional subsample restrictions appear in the robustness battery. The "no tribal" restriction drops counties whose AIAN share exceeds 10% in any panel year; this is a stricter cut than the high-AIAN split and removes 1,234 county-year observations. The "no top metro" restriction drops counties in the top within-year population quartile and serves as a direct test of whether the headline result is driven by the largest metropolitan areas.

6 Results and Interpretation

This section reports the empirical results and discusses their interpretation as the evidence unfolds. Throughout, the outcome is the inverse hyperbolic sine of real per-capita competitive grant obligations, and standard errors are clustered at the county level unless otherwise noted.

6.1 Headline: PDV Rank and Competitive Grants

As shown in Table 3, the baseline two-way fixed effects regression of equation (10) produces a coefficient on $\text{PDV}_{i,t}^{\text{pct}}$ of approximately 0.0022, statistically significant at the 5% level, and stable in magnitude across specifications that progressively add fixed effects and controls. Moving from county and year fixed effects only (column 1) to county and state-by-year fixed effects with the full control vector (column 4), the estimated coefficient remains essentially unchanged. The headline specification implies that a 10-percentile improvement in a county's within-year PDV rank is associated with approximately a 2.2% increase in real per-capita competitive grant obligations. A 25-percentile-point improvement (for example, a movement from the median to the 75th percentile) implies a 5.5% increase; a movement across the full interquartile range, from the 25th to the 75th percentile, implies approximately an 11% increase. The lagged-PDV specification in column 5 yields a slightly smaller coefficient of 0.0018, marginally significant at the 10% level, indicating that the relationship operates with a timing horizon of at most one year and is somewhat stronger contemporaneously.

Translating this proportional effect into dollar magnitudes requires specifying the baseline level of per-capita receipt at which the elasticity is evaluated, because the inverse hyperbolic sine transformation maps proportional changes back to dollars only locally. Evaluated at the panel-wide unconditional mean of \$54 per capita, the 11% interquartile-range effect implies approximately \$6 per capita in additional competitive grant capture, or approximately \$300,000 annually for a county of fifty thousand residents. Evaluated at the higher conditional mean of \$122 per capita that obtains among the 44% of county-years with positive receipts, the same 11% effect implies approximately \$13 per capita, or approximately \$650,000 annually for the same hypothetical county. Total competitive grant volume in the United States averages approximately \$25 billion per year over the panel period; a reallocation driven by visibility improvements at this magnitude is meaningful for the counties that move in the within-year PDV distribution. The headline result is consistent with Proposition 1: among visible counties competing for the federal agency's budget, higher visibility translates into higher allocation.

Figure 7 reports the bin-scatter representation of the headline relationship. The horizontal axis groups counties into fifty within-year percentile bins of PDV; the vertical axis plots the mean of the IHS-transformed real per-capita competitive grant obligation in each bin, after partialling out log population and state-by-year fixed effects. The bin-scatter is approximately linear across the bulk of the distribution, providing visual reassurance that the linear specification in equation (10) captures the empirical relationship without imposing strong functional form assumptions.

6.2 Long-Difference Robustness

The long-difference specification, which exploits only the 2022-minus-2012 first difference for each county, produces a coefficient on $\Delta\text{PDV}_i^{\text{pct}}$ of 0.0034, marginally significant at the 10% level, as shown in Table 4. Adding the standard time-varying controls in difference form sharpens the estimate to 0.0034 with a standard error of 0.0019, again marginally significant. The within-county nature of the long-difference design uses entirely different identifying variation than the year-on-year movements exploited by the TWFE specification, and the consistency between the two coefficients, 0.0022 in TWFE and 0.0034 in long difference, strengthens confidence that the result reflects genuine within-county movement of PDV rank rather than a spurious cross-sectional correlation.

6.3 Rank Versus Level

A direct comparison of three alternative transformations of PDV yields one of the paper's most striking results. In Table 5, the continuous 0–4 raw composite produces a coefficient of -0.0848, statistically indistinguishable from zero but pointed in the wrong direction. The within-year z-score produces a coefficient of -0.0042, also statistically indistinguishable from zero and again with the wrong sign. Only the within-year percentile rank produces a significant positive coefficient. The level-based specifications do not merely fail to detect a relationship; they detect a small negative relationship that disappears once the variation is reframed as a within-year rank.

This finding is the empirical realization of Proposition 3: the federal allocation apparatus operates on within-year rank rather than on absolute level. A county that improves its raw PDV by one point while the national mean rises by the same amount gains no allocative advantage; only a county that improves its rank captures the implied benefit. The pattern aligns with a broader institutional development. The federal allocation apparatus has increasingly adopted rank-based screening tools: the Climate and Economic Justice Screening Tool used to designate Justice40 disadvantaged communities, the USDA Persistent Poverty Counties designation, the IRA Energy Communities provisions, and HUD's distress indices all use percentile thresholds rather than absolute indicators. A county that improves its absolute PDV without improving its rank gains no allocative advantage in any of these systems. The raw-PDV null with the percentile-PDV positive is the empirical realization of this institutional fact, and the finding implies that future empirical work on visibility-mediated allocation should specify the regressor in rank form: absolute visibility measures may not just miss relationships that rank-based specifications detect but may detect spurious null or wrong-signed associations.

6.4 Mechanism: Sub-Index Decomposition

The decomposition of the headline effect into the three sub-indices of PDV identifies a single operative dimension. In Table 6, when the Coverage, Resolution, and Usability percentile ranks enter the regression individually, only Coverage produces a positive and significant coefficient (0.0022, significant at the 1% level); Resolution and Usability are essentially zero (0.0003 and -0.0002 respectively). When all three enter jointly, the Coverage sub-index dominates: its coefficient rises to 0.0027 and remains highly significant, while Resolution and Usability remain individually insignificant. The long-difference specification in Table 4 reinforces this finding: the within-county increase in Coverage percentile predicts grant capture (ΔCoverage coefficient of 0.0036, significant at the 5% level), while changes in Resolution and Usability percentiles do not.

This pattern identifies the operating mechanism with unusual sharpness. The dominance of the Coverage sub-index — the extensive margin of visibility, the binary fact of being measurable in a domain — aligns precisely with the threshold interpretation of visibility in Proposition 2. What matters for allocation is whether the county crosses the visibility threshold and enters the considered set; refinement of existing data (Resolution) and improvements in technical accessibility (Usability) do not detectably affect allocation once Coverage itself is controlled for. The implication is direct: the relevant policy lever for reducing visibility-mediated allocation inequality is investment in the extensive margin — making previously invisible counties measurable in any form — rather than refinements of already-visible data. The contemporary debates about federal data modernization (the Census Bureau Data Modernization Initiative, the CDC Public Health Data Modernization Initiative, the FCC Broadband Data Collection rebuild) have emphasized technical accessibility and granularity, but the analysis here suggests that the binding constraint for allocation is whether the underlying data exist at all for a given county. State open-data laws affect not only transparency and civic participation — the channels usually emphasized — but also operate through a downstream allocative channel that reshapes federal resource flow.

6.5 Domain Decomposition

In Table 7, eight PDV domain scores enter the regression individually with markedly different coefficients. Schools, Environment, and Health produce the largest and most significant positive coefficients (0.0062, 0.0029, and 0.0025 respectively, all significant at the 1% level). The Broadband and Business/Labor domains produce coefficients indistinguishable from zero — in fact effectively zero given the limited cross-county variation in those domain scores. The remaining domains (Housing, Transportation, Local Finance) produce small

and statistically insignificant coefficients.

The pattern maps cleanly to the program composition of the competitive grants panel. NIH research grants, which dominate the health-related dollar volume in the competitive grants panel, draw on public health surveillance data captured by the Health domain (CDC PLACES, HRSA Area Health Resources Files). EPA Brownfields and Great Lakes grants, dominant in the environmental component of the panel, draw on EJScreen, TRI, and AQS measurements captured by the Environment domain. Most strikingly, Promise Neighborhoods and ED Institute of Education Sciences research grants, both heavy in the panel, draw on NCES school-level and district-level data captured by the Schools domain. When all eight domain percentiles enter the regression jointly, Schools (0.0061), Health (0.0023), and Environment (0.0023) retain individual significance, consistent with these being the domains where the underlying funded programs are largest. The joint F-test of all eight coefficients equal to zero is rejected. The near-zero coefficients on Broadband and Business/Labor are not a puzzle: the underlying domain scores have essentially no cross-county variation, so these domains contribute almost no identifying variance.

6.6 Causal Identification

The instrumental variables results in Table 8 provide consistent evidence that the within-county TWFE relationship reflects a causal effect.

The first-stage regression of PDV rank on the state open-data law indicator yields a coefficient of 2.0028, highly significant. The two-instrument first stage including both the open-data law and the Chief Data Officer indicator yields 2.2172 for the open-data law and 1.5199 for the CDO position, both highly significant. The Kleibergen-Paap weak-IV F-statistic is 22.37 in the single-instrument specification and 15.13 in the two-instrument specification, both comfortably above conventional weak-IV thresholds.

The single-instrument two-stage least squares estimate of the visibility effect is 0.0267, more than ten times larger than the OLS estimate. The two-instrument estimate is 0.0290, comparable in magnitude. Adding state-specific linear time trends to the IV specification yields a still-larger coefficient of 0.1855, suggesting that the absorbed trends were attenuating the causal estimate downward. The Bartik shift-share instrument has a strong first stage (coefficient 38.4272 with weak-IV F-statistic of 54.26) and yields a second-stage coefficient of 0.0775, all significant at the 1% level.

Across all four IV specifications, the estimated visibility effect is positive, statistically significant, and substantially larger than the OLS estimate. The pattern is consistent with reverse causality biasing the OLS estimate downward: federal grants come with reporting

requirements that drag visibility upward, so the unbiased causal effect of visibility on grants must be larger than the contemporaneously estimated within-county relationship suggests. The IV results confirm that the headline finding reflects a causal effect of PDV rank on competitive grant capture rather than reverse causation or unobserved confounding.

6.7 Placebo Battery: Specificity to Competitive Grants

The placebo battery test in Table 9 provides clean evidence for the visibility-screening mechanism. The coefficient on PDV percentile rank is positive and significant (0.0022) only for competitive grants. For broad-classified discretionary grants (which include quasi-formula programs excluded from the competitive bucket), the coefficient is 0.0006 and statistically insignificant. For formula grants the coefficient is 0.0005, insignificant. For direct entitlement payments the coefficient is -0.0001, indistinguishable from zero. For direct loans the coefficient is 0.0009, insignificant. For total federal assistance the coefficient is essentially zero.

This pattern is the empirical realization of Proposition 4: the visibility effect operates only where federal officer discretion operates. The null effect on formula grants confirms that the headline result is not driven by PDV proxying for generic state capacity or political connections; if it were, formula grants would also load on PDV. The null effect on direct entitlement payments confirms that the result is not driven by PDV proxying for population demographics; direct payments flow to individuals based on eligibility regardless of county data infrastructure. The null effect on direct loans confirms that the result is not driven by PDV proxying for creditworthiness; loan volume does not depend on county documentation in the same way that grant applications do. These nulls are not failures of the model but its boundary conditions: formula-allocated programs by definition do not respond to applicant visibility, direct entitlement payments flow to eligible individuals regardless of county documentation, and direct loans are extended by lenders who evaluate creditworthiness on different criteria than federal officers evaluate grant applications. The asymmetric empirical pattern across grant types confirms that the visibility mechanism operates exactly where the theory predicts it should and not where it should not.

6.8 Heterogeneity

The heterogeneity patterns reported in Tables 10, 11, and 12 reveal a more nuanced picture than a uniform visibility effect would suggest.

By population quartile. In Tables 10 , counties are split into quartiles within each fiscal year using Census PEP annual population, so the cutoffs are the 25th, 50th, and 75th within-year percentiles. In the smallest-quartile counties (Q1), the coefficient is -0.0005 and insignificant. In second-quartile counties (Q2), the coefficient is 0.0040 and statistically significant at the 5% level. In third-quartile and fourth-quartile counties (Q3 and Q4), the coefficients are 0.0030 and 0.0010 respectively, both statistically insignificant. The effect is therefore concentrated in mid-sized counties — not in the smallest counties where competition for grants is limited and visibility may not matter, and not in the largest counties where visibility is uniformly high and rank movements are mechanically constrained.

This pattern is consistent with the threshold interpretation of the model developed in Section 3. The marginal effect of visibility on allocation is greatest for counties near the visibility threshold $\underline{v}$ — counties whose small visibility improvements move them across the boundary between being excluded and being considered. Very small counties lie well below the threshold even after improvement; very large counties lie well above it. The Q2 counties (and to a lesser extent Q3) are the population layer at which marginal visibility improvements have the largest allocative consequences. The empirical pattern is therefore not a refutation of the model but a confirmation of its threshold structure operating non-uniformly across the size distribution.

By era. Pre-DATA-Act years are 2012–2016, before the full compliance deadline of the Digital Accountability and Transparency Act in May 2017; post-DATA-Act years are 2017–2022. Pre-COVID years are 2012–2019; post-COVID years are 2020–2022. The PDV effect is concentrated in the earlier portion of the panel. In pre-DATA Act years, the coefficient is 0.0028 and statistically significant at the 5% level. In post-DATA Act years, the coefficient is 0.0002 and statistically insignificant. Splitting at the COVID divide produces a similar pattern: pre-COVID coefficient of 0.0036 (highly significant at the 1% level) versus post-COVID coefficient of 0.0023 (insignificant).

This is one of the more interesting findings of the analysis and warrants careful interpretation. Two complementary explanations are consistent with the pattern. The first is that the DATA Act of 2014, fully effective in 2017, substantially standardized federal data reporting and reduced cross-county variation in how counties appeared in federal datasets. The Act mandated machine-readable reporting of federal financial assistance and required agencies to report transactions in standardized form. To the extent that this standardization narrowed the cross-county dispersion of visibility, the marginal effect of any individual county's visibility movement would shrink mechanically. The threshold $\underline{v}$ in the model may have effectively risen as the floor of acceptable documentation rose, reducing the share of

the population in the marginal range where visibility movements matter most. The second is that the post-2020 period saw substantial pandemic-era federal allocation flow through formulaic channels — the Coronavirus Relief Fund, ARPA State and Local Fiscal Recovery Funds, the American Rescue Plan capital projects fund — that bypassed the competitive selection mechanism through which visibility operates. To the extent that these emergency programs represented a larger share of total federal allocation in the post-COVID years, the competitive channel through which the visibility effect operates was a smaller share of the underlying allocation environment.

Both explanations imply that the visibility-allocation relationship is policy-environment-dependent rather than fixed. The relevance of visibility for federal allocation depends on the institutional features of the allocation apparatus, and changes to that apparatus — through standardization mandates, expansion of formulaic allocation, or the introduction of new screening tools — can alter the magnitude of the relationship. This is itself a finding consistent with Scott's framework: legibility's relationship to state action is not a fixed coefficient but a contingent product of institutional design.

By racial composition. In Table 12, counties are classified as high-Black if their Census PEP Black share exceeds 15%, high-Hispanic if their Hispanic share exceeds 15%, and high-AIAN if their American Indian / Alaska Native share exceeds 5%. The corresponding low-share subsamples are the complements — counties below each cutoff. The pattern of statistical significance across these six subsamples is uneven. The estimated coefficient is positive and statistically significant in counties with low shares of each minority group: 0.0022 in low-Black counties (significant at the 5% level), 0.0019 in low-Hispanic counties (significant at the 10% level), and 0.0021 in low-AIAN counties (significant at the 5% level). In the high-share subsamples, the coefficient is positive in every case but is not statistically distinguishable from zero at conventional levels: 0.0025 in high-Black counties (SE 0.0025), 0.0038 in high-Hispanic counties (SE 0.0030), and 0.0042 in high-AIAN counties (SE 0.0046). The high-share subsamples are also markedly smaller (between 2,164 and 6,874 county-years, against 27,245 to 31,917 in the corresponding low-share subsamples).

The group equality test shows that there is no significant difference in the coefficient between two groups. The visibility-grants relationship is therefore not narrowly confined to majority-white counties. The pattern is consistent with the visibility-allocation mechanism operating broadly across the racial composition of the panel rather than being driven by demographic targeting in particular subsets of the country.

6.9 Robustness

The robustness battery in Tables 13 and 14 confirms that the headline finding is not fragile to specification choices. Replacing the IHS-transformed outcome with $\log(1 + y)$, the untransformed level, the binary extensive-margin indicator "any competitive grant received," the within-year national share, the log allocation ratio, or the count of distinct programs received all yield positive and significant coefficients. The extensive-margin specification — whether the county received any competitive grant in the fiscal year — yields a coefficient of 0.0006, significant at the 1% level, providing a particularly clean realization of the threshold mechanism in Proposition 2. The count-of-programs specification yields 0.0018, also significant at the 1% level.

Subsample restrictions produce consistent results with two informative variations. The "no tribal" restriction drops counties whose Census PEP AIAN share exceeds 10% in any panel year — a stricter cut than the high-AIAN heterogeneity split, removing 1,234 county-year observations from the panel. This restriction yields a coefficient of 0.0021, essentially unchanged from the headline. The "no top metro" restriction drops counties in the top within-year population quartile, restricting attention to roughly the bottom 75% of the size distribution; it yields a coefficient of 0.0025, modestly *larger* than the full sample headline. This pattern is directly consistent with the population quartile heterogeneity above: removing the Q4 counties that show the weakest visibility effect produces a more precisely estimated and slightly larger average effect among the remaining counties. The result confirms that the headline is not driven by the largest metropolitan areas and indeed strengthens when those areas are excluded.

7 Conclusion

This paper develops and tests the proposition that the geographic distribution of administrative legibility shapes the geographic distribution of federal resources. Using a new county-year measure of public data visibility spanning eight substantive domains and sixteen federal data sources, linked to a hand-curated panel of competitive federal grant programs, we find that a county's within-year percentile rank in the national visibility distribution predicts its capture of competitive grant funding. A 10-percentile improvement is associated with approximately a 2.2% increase in real per-capita competitive grant obligations. The effect is driven by the extensive margin of visibility, whether a county clears the threshold of measurability at all, and is entirely absent in the placebo categories of formula grants, direct payments, and direct loans, where administrative screening does not operate. Instrumental

variables estimates confirm a positive causal effect.

The results provide quantitative evidence for Scott's classical legibility hypothesis in the contemporary U.S. federal allocation context. They suggest that investments in county-level data infrastructure have meaningful allocative consequences and that the geographic distribution of visibility itself constitutes a previously unmeasured dimension of measurement inequality. The findings have implications for the design of federal data modernization initiatives, the policy framing of state open-data laws, and the empirical analysis of federal grant geography. The era pattern documented in the heterogeneity analysis suggests further that the visibility- allocation relationship is contingent on the institutional features of the federal allocation apparatus and is not a fixed parameter of the U.S. political economy.

The findings open several questions for future work. How does the visibility–allocation channel interact with the algorithmic targeting tools that increasingly mediate federal disbursement? What role does visibility play in non-federal allocation domains — foundation philanthropy, municipal bond markets, local news coverage? How does the contemporary visibility distribution map to historical patterns of statistical undercount?

References

Acemoglu, D., Cheema, A., Khwaja, A. I., and Robinson, J. A. (2016). Trust in state and nonstate actors: Evidence from dispute resolution in pakistan. Working Paper 22184, NBER.

Acemoglu, D., García-Jimeno, C., and Robinson, J. A. (2015). State capacity and economic development: A network approach. *American Economic Review*, 105(8):2364–2409.

Adelino, M., Cunha, I., and Ferreira, M. (2017). The economic effects of public financing: Evidence from municipal bond ratings recalibration. *Review of Financial Studies*, 30(9):3223–3268.

Bellemare, M. F. and Wichman, C. J. (2020). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics*, 82(1):50–61.

Besley, T. and Persson, T. (2009). The origins of state capacity: Property rights, taxation, and politics. *American Economic Review*, 99(4):1218–1244.

Caplin, A. and Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203.

Cellini, S. R., Ferreira, F., and Rothstein, J. (2010). The value of school facility investments: Evidence from a dynamic regression discontinuity design. *Quarterly Journal of Economics*, 125(1):215–261.

Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *Quarterly Journal of Economics*, 129(4):1553–1623.

Chodorow-Reich, G. (2019). Geographic cross-sectional fiscal spending multipliers: What have we learned? *American Economic Journal: Economic Policy*, 11(2):1–34.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624.

Harrison, T. M., Guerrero, S., Burke, G. B., Cook, M., Cresswell, A., Helbig, N., Hrdinová, J., and Pardo, T. (2012). Open government and e-government: Democratic challenges from a public value perspective. *Information Polity*, 17(2):83–97.

Hogan, H. (2003). Accuracy and coverage evaluation: Theory and application. Technical report, U.S. Census Bureau, Statistical Research Division.

Ruijer, E., Grimmelikhuijsen, S., van den Berg, M., and Meijer, A. (2020). Open data work: Understanding open data usage from a practice lens. *International Review of Administrative Sciences*, 86(1):3–19.

Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Suárez Serrato, J. C. and Wingender, P. (2016). Estimating local fiscal multipliers. Working Paper 22425, NBER.

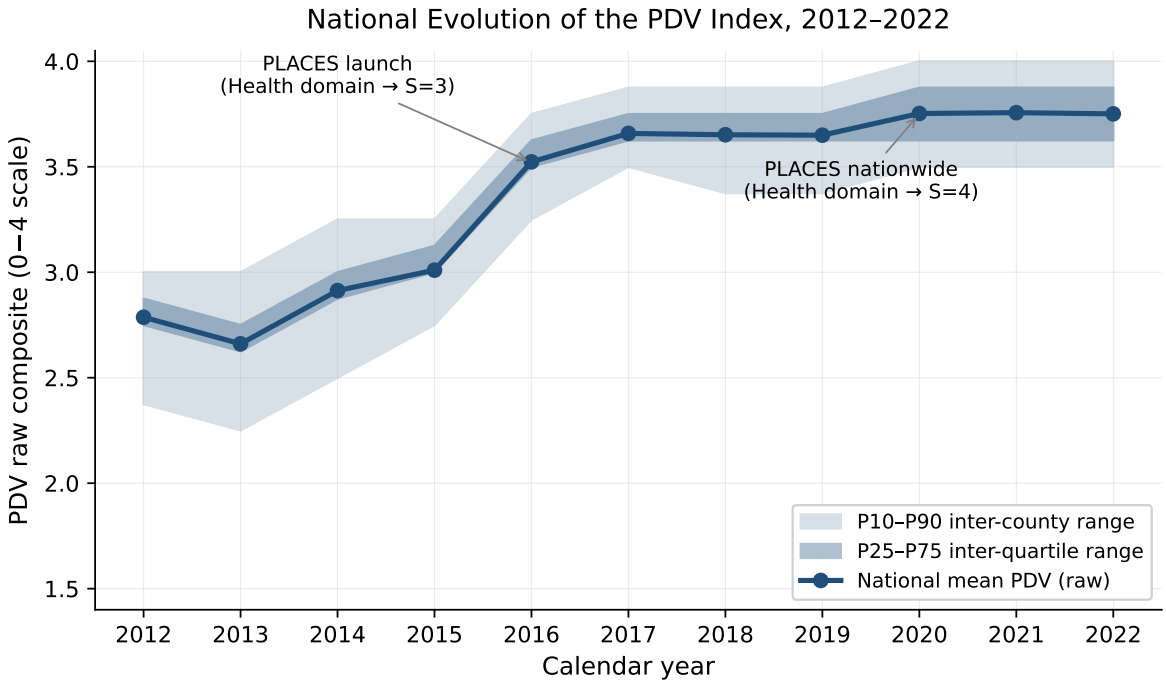

Figure 1: National evolution of the Public Data Visibility (PDV) composite index, 2012–2022. The solid line plots the cross-county mean of the raw 0–4 composite in each fiscal year. The dark band spans the 25th–75th within-year percentiles; the light band spans the 10th–90th. Visible step changes correspond to the launch of new federal data sources (CDC PLACES in 2016, nationwide PLACES expansion in 2020). The within-year dispersion of PDV narrows over the panel period as national data infrastructure expands.

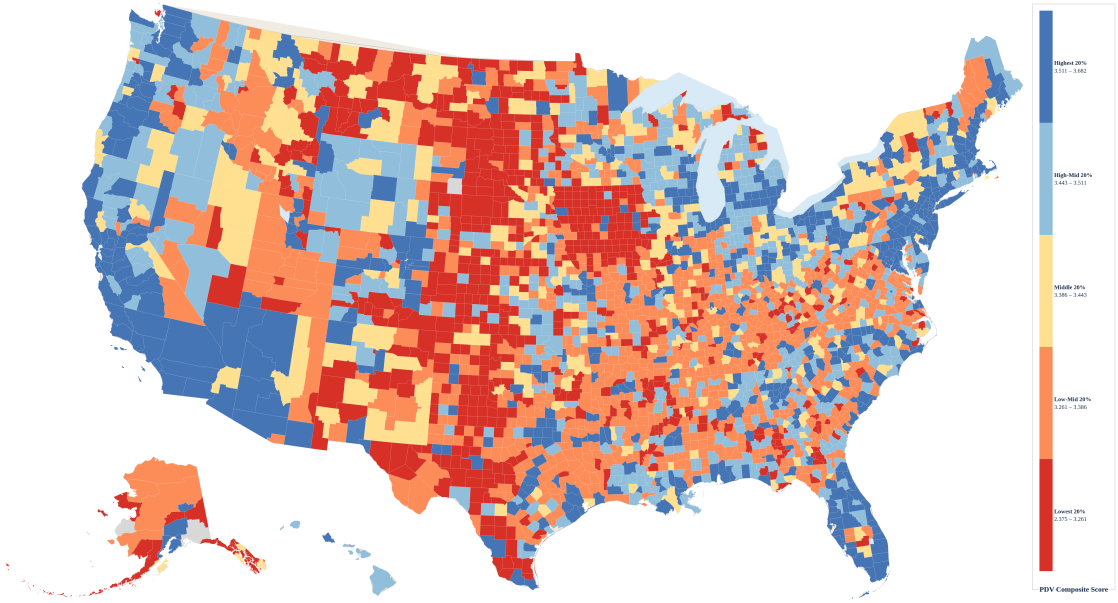

Figure 2: Public Data Visibility: Composite Score (County Panel Averages, 2012-2022). Each county's average raw PDV composite score over the full 2012-2022 panel. The composite is the simple mean of eight domain scores(Health,Environment,Broadband, Housing, Transportation, Schools, Local Finance, Business & Labor), each from {0, 1, 2, 3, 4}. Counties are assigned to quintile bins; colour runs from red

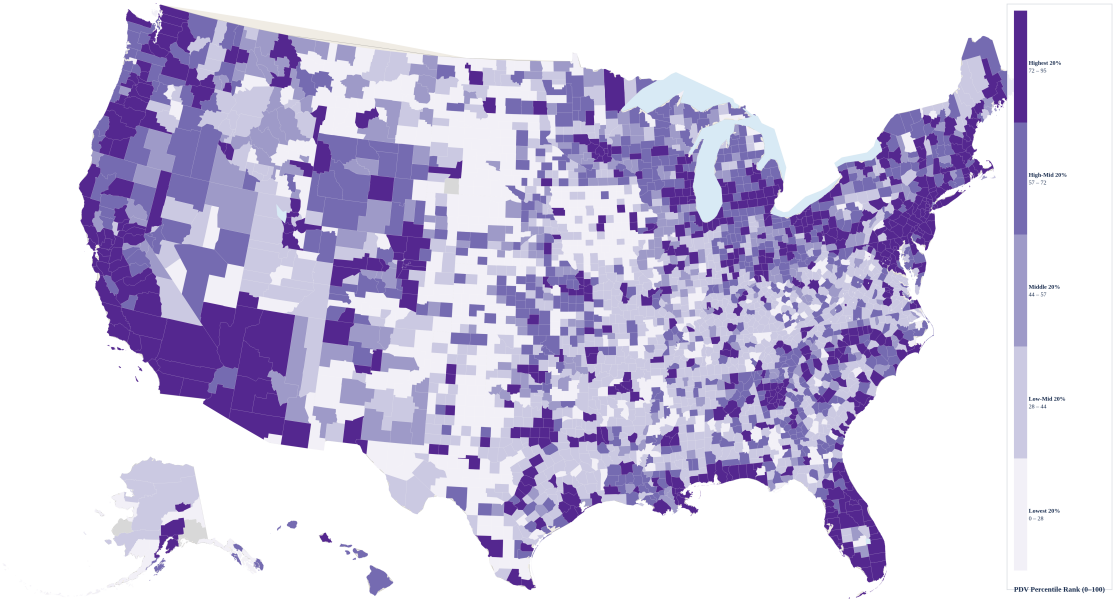

Figure 3: Public Data Visibility: Within-Year Percentile Rank (County Panel Averages,2012-2022). Each county's PDV score re-expressed as a within-year percentile rank (0-100) relative to all other county-year observations in the same calendar year, then averaged across years. Removes secular trends; isolates persistent cross-sectional differences in relative visibility. Colour runs from light lavender (low-ranked) to deep violet (high-ranked).

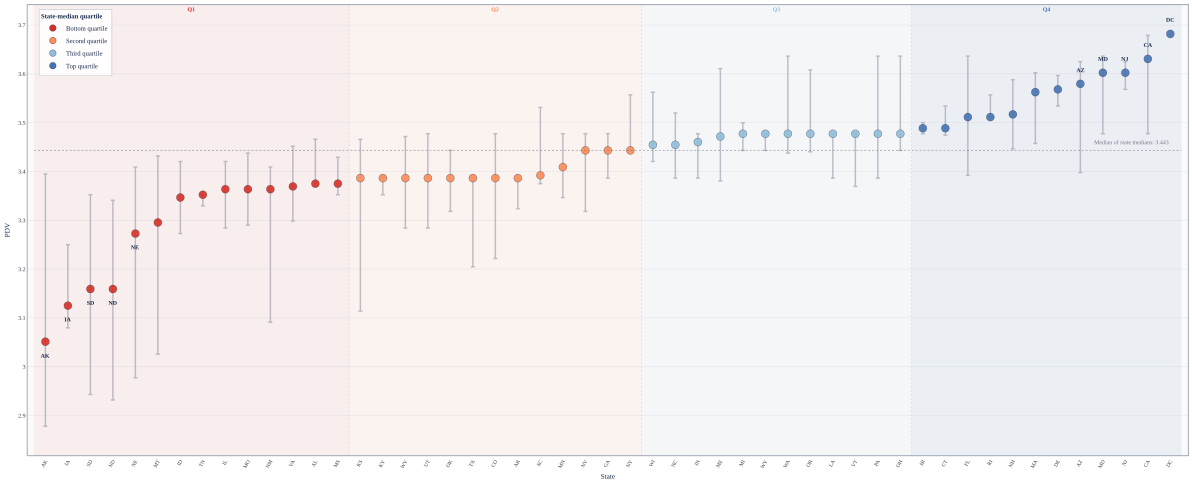

Figure 4: State-Level PDV Distribution (County Panel Averages, 2012- 2022). Each point is one state's median county PDV score (panel averages 2012-2022), sorted lowest to highest along the horizontal axis. Vertical bars show the within-state interquartile range (IQR) of county PDV averages. Colour denotes national-rank quartile: red (bottom) through blue (top). The five lowest (AK, IA, SD, ND, NE) and five highest (AZ, MD, NJ, CA, DC) states are labelled. The dotted line marks the median of state medians (3.443). The substantial cross-state variation in PDV supports the use of state open-data laws as an instrumental variable for PDV in causal identification.

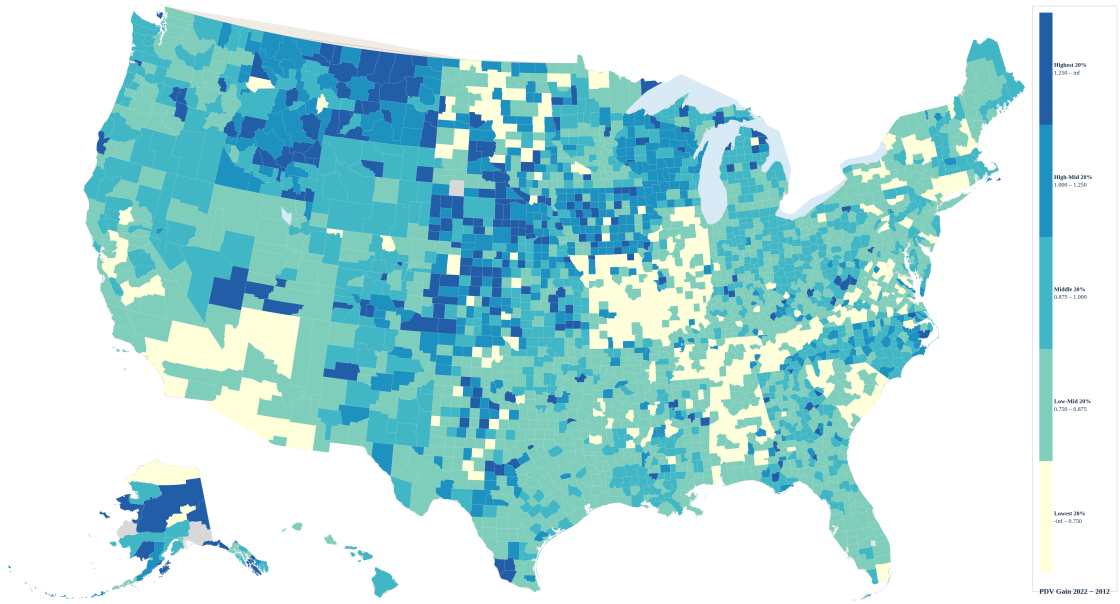

Figure 5: Ten-Year Change in Public Data Visibility by County, 2012-2022. The 2012-to-2022 gain in the raw PDV composite ($\Delta PDV_i = PDV_{i,2022} - PDV_{i,2012}$) for 3,143 counties with observations at both endpoints. Every county recorded a strictly positive gain: range +0.375 to +1.875, national median +0.875. Because ΔPDV is constrained to multiples of 0.125, bins use natural value-based breakpoints (+0.750 / +0.875 / +1.000 / +1.250). Colour runs from light yellow-green (smallest gain) to deep navy blue (largest gain).

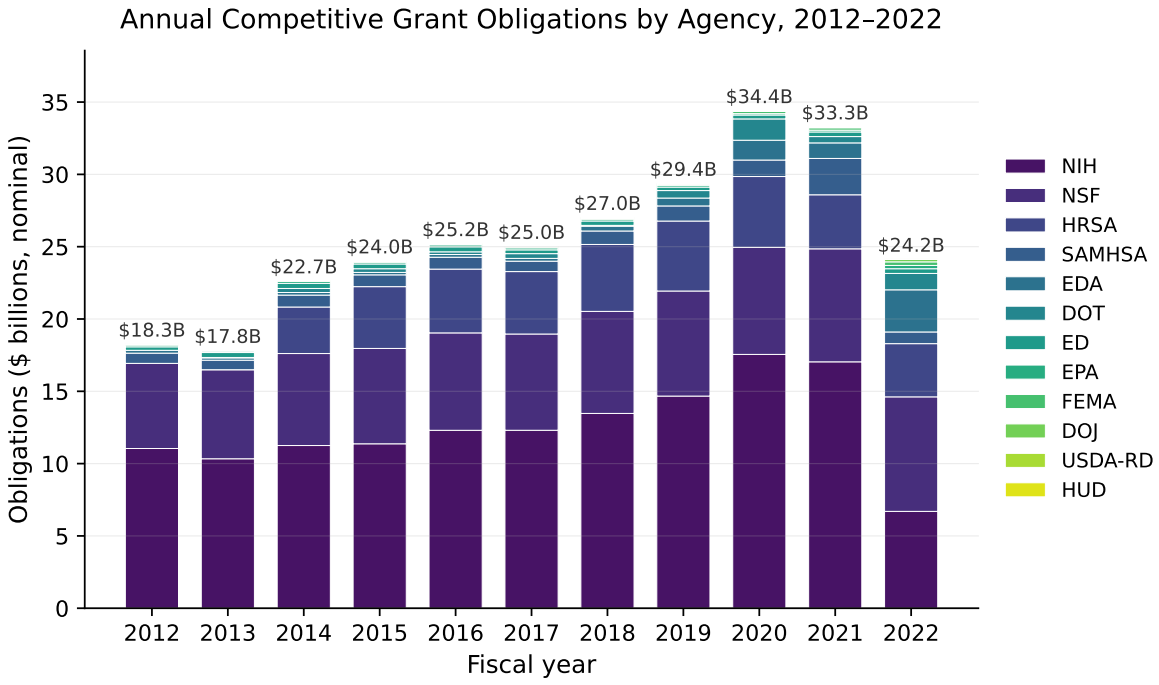

Figure 6: Annual competitive federal grant obligations by federal agency, 2012–2022. Bars stack contributions from the eleven federal agencies that contribute the 34 programs in the competitive grants panel. Annual totals (in billions of nominal dollars) are labelled above each bar. Biomedical research funding (NIH) dominates throughout the panel period; the 2020 peak reflects pandemic-era discretionary supplements administered through competitive review.

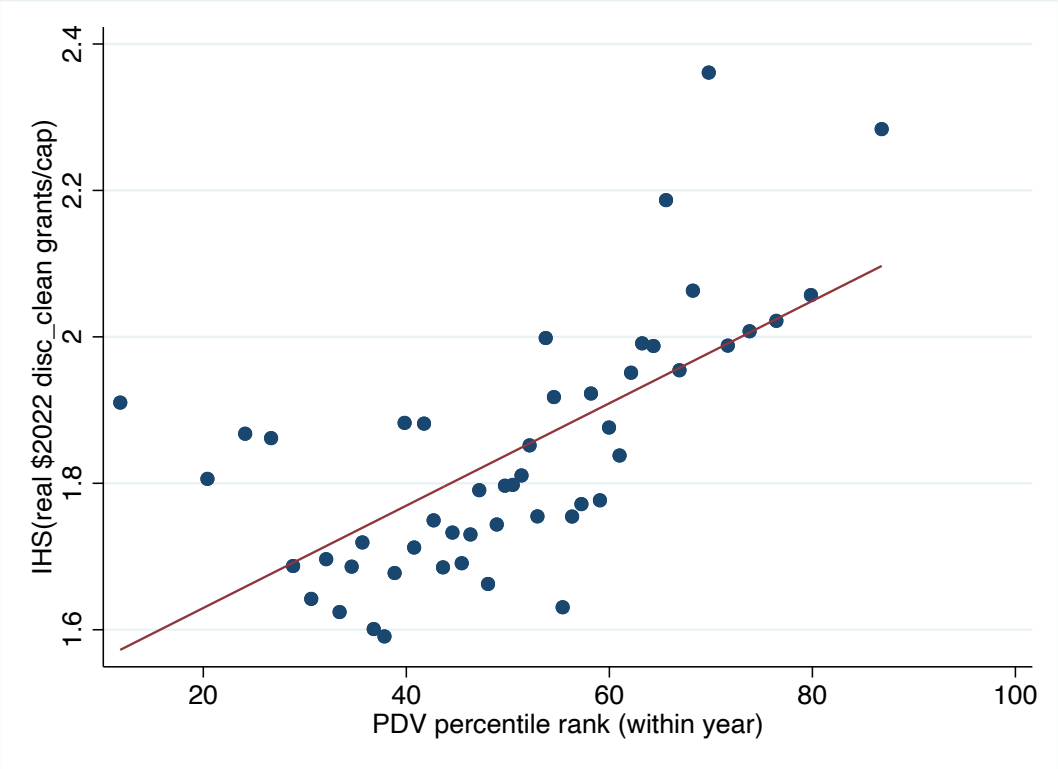

Figure 7: Bin-scatter of the headline relationship between PDV percentile rank and the inverse-hyperbolic-sine of real per-capita competitive grant obligations. Each point represents the mean of the outcome variable for one of fifty within-year PDV percentile bins, after partialling out log population and state-by-year fixed effects. The slope corresponds to the headline coefficient $\beta \approx 0.0022$ from Table**, column 4.

Table 1: National Evolution of the PDV Composite and Sub-Indices, 2012–2022

Year	PDV raw (0–4)	Coverage (0–1)	Resolution (0–4)	Usability (0–4)
2012	2.79	0.72	2.07	3.27
2013	2.66	0.72	1.95	3.26
2014	2.91	0.85	2.20	3.26
2015	3.01	0.87	2.42	3.34
2016	3.52	0.99	2.68	3.84
2017	3.66	0.99	2.80	3.96
2018	3.65	0.99	2.68	3.97
2019	3.65	0.99	2.69	3.97
2020	3.75	0.99	2.80	3.97
2021	3.76	0.99	2.80	3.97
2022	3.75	0.99	3.05	3.97
Pooled mean	3.37	0.92	2.56	3.71
Pooled SD	0.47	0.11	0.36	0.36

Table 2: Summary statistics for the analysis panel, 2012–2022

Variable	count	mean	sd	min	p25	p50	p75	max
Outcome: Public Data Visibility index								
PDV_raw	34533	3.37	0.47	1.62	3.00	3.50	3.75	4.00
PDV_z	34533	0.00	1.00	-4.44	-0.16	0.30	0.60	1.66
PDV_pct	34533	50.04	27.68	0.03	36.88	54.14	71.13	96.52
PDV_coverage_pct	34533	50.03	14.62	1.73	53.09	53.32	53.63	59.20
PDV_resolution_pct	34533	50.02	25.59	1.73	27.65	51.67	74.47	95.70
PDV_usability_pct	34533	50.03	14.62	1.73	53.09	53.32	53.64	59.20
Domain sub-scores								
score_health_pct	34533	50.03	9.36	0.16	44.62	50.02	50.02	94.62
score_environment_pct	34533	50.03	16.96	6.60	56.60	56.63	56.71	56.73
score_broadband_pct	34533	50.04	0.04	50.02	50.02	50.02	50.05	50.14
score_housing_pct	34533	50.06	16.36	5.77	56.09	56.14	56.22	56.35
score_transportation_pct	34533	50.01	24.63	3.01	28.54	46.93	75.09	92.64
score_schools_pct	34533	50.03	15.96	1.67	51.72	56.08	57.65	62.12
score_local_finance_pct	34533	50.01	9.25	2.08	50.02	50.02	53.75	55.03
score_business_labor_pct	34533	49.99	20.20	39.72	39.72	39.72	39.73	89.73
Outcome: competitive federal grants								
grants_disc_clean_pc_real22	34533	53.83	244.04	-8289.29	0.00	0.00	27.88	19268.68
n_programs_disc_clean	34533	2.11	4.73	0.00	0.00	0.00	2.00	30.00
any_disc_clean	34533	0.44	0.50	0.00	0.00	0.00	1.00	1.00
Controls								
pop	34533	103020	329330	44	10903	25727	67771	10105708
unemp_rate	34458	5.47	2.42	1.10	3.70	4.90	6.70	27.40
poverty_pct	34516	15.58	6.23	2.60	11.10	14.50	18.90	56.70
median_hh_income_real22	34516	61053	15947	25896	50471	58585	68189	181268
share_black	34533	0.09	0.14	0.00	0.01	0.02	0.11	0.87
share_hisp	34533	0.09	0.14	0.00	0.02	0.04	0.10	0.98
share_65plus	34533	0.19	0.05	0.00	0.16	0.18	0.21	0.58
dem_voteshare_pres	34234	0.33	0.16	0.02	0.21	0.31	0.43	0.91
n_disasters	34533	0.56	0.98	0.00	0.00	0.00	1.00	10.00

Notes. County×fiscal-year panel covering 3,144 U.S. counties and county-equivalents over fiscal years 2012–2022; the analysis sample is 34,533 observations after dropping county-years with missing population or invalid FIPS codes. Variables in the upper panels are constructed for this paper: `PDV_raw` is the composite Public Data Visibility score on a 0–4 ordinal scale; `PDV_z` is the within-year z-score; `PDV_pct` and the eight domain-level `_pct` variables are within-year percentile ranks (0–100). The composite is the simple average of eight domain scores (health, environment, broadband, housing, transportation, schools, local finance, and business/labor) and decomposes into three sub-indices: Coverage (the extensive margin of visibility), Resolution (spatial granularity), and Usability (technical accessibility). The competitive-grants outcome `grants_disc_clean_pc_real22` aggregates obligations from 34 hand-curated peer-reviewed and discretionary federal grant programs across eleven agencies (full program list in Appendix B), reported in real 2022 dollars per capita; `n_programs_disc_clean` is the count of those 34 programs from which the county received any obligation in the fiscal year; and `any_disc_clean` is the extensive-margin indicator. Controls are drawn from Census PEP, BLS LAUS, Census SAIPE, MIT MEDSL, and FEMA OpenFEMA. County identifiers are 5-digit zero-padded FIPS codes.

Table 3: Headline two-way fixed-effects estimates: visibility rank and competitive federal grant capture

	(1) County+Yr FE	(2) County+SY FE	(3) + Short ctrl	(4) + Full ctrl	(5) PDV lagged
PDV_pct	0.0023**	0.0023**	0.0023**	0.0022**	
	(0.0009)	(0.0010)	(0.0010)	(0.0010)	
log_pop			-0.0506	0.1633	0.2587
			(0.3561)	(0.3568)	(0.3608)
unemp_rate			-0.0514***	-0.0514***	-0.0347**
			(0.0162)	(0.0163)	(0.0161)
poverty_pct			0.0010	-0.0019	0.0010
			(0.0069)	(0.0070)	(0.0070)
dem_voteshare_pres			-0.1016	-0.1693	0.0232
			(0.3244)	(0.3210)	(0.3099)
share_black				0.5648	0.1538
				(2.4005)	(2.4274)
share_hisp				0.8311	-0.1983
				(2.0447)	(1.9075)
share_65plus				5.6027***	4.5745**
				(1.8575)	(1.8639)
median_inc_10k				-0.0249	-0.0154
				(0.0322)	(0.0320)
n_disasters				0.0252	0.0237
				(0.0165)	(0.0168)
PDV_pct_l1					0.0018*
					(0.0010)
_cons	1.7261***	1.7246***	2.5167	-0.6542	-1.4844
	(0.0454)	(0.0498)	(3.6420)	(3.6122)	(3.6622)
Observations	34533	34522	34162	34162	31057
Within R-sq	0.000	0.000	0.001	0.002	0.001

Notes. The dependent variable is the inverse hyperbolic sine of real-2022-dollar per-capita competitive federal grant obligations from the 34-program panel described in Appendix B. The regressor of interest, `PDV_pct`, is the within-year percentile rank (0–100) of the county's Public Data Visibility composite. Column 1 absorbs county and calendar-year fixed effects; columns 2–5 absorb county and state-by-year fixed effects. Column 3 adds short controls (`log_pop`, `unemp_rate`, `poverty_pct`, `dem_voteshare_pres`); column 4 adds the full control vector (the short controls plus `share_black`, `share_hisp`, `share_65plus`, `median_inc_10k` in $10,000 units, and `n_disasters`). Column 5 replaces contemporaneous `PDV_pct` with its one-year lag (`PDV_pct_l1`); the smaller sample reflects the loss of fiscal year 2012. Standard errors clustered at the county level appear in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Long-difference (2022 vs. 2012) within-county estimates with sub-index decomposition

	(1) PDV_pct	(2) + ctrl	(3) Coverage_pct	(4) Resolution_pct	(5) Usability_pct
d_PDV_pct	0.0032*	0.0034*			
	(0.0019)	(0.0019)			
d_PDV_coverage_pct			0.0036**		
			(0.0016)		
d_PDV_resolution_pct				0.0009	
				(0.0016)	
d_PDV_usability_pct					-0.0030
					(0.0022)
Observations	3133	3105	3105	3105	3105
R-sq	0.001	0.013	0.012	0.012	0.012

Notes. The sample collapses the panel to two endpoints, fiscal years 2012 and 2022, and forms first differences at the county level; each observation is one county. The dependent variable (Δy) is the 2022-minus-2012 change in the inverse hyperbolic sine of real-2022-dollar per-capita competitive grants. Each row reports the coefficient on the corresponding first-differenced regressor: change in PDV percentile rank (columns 1–2), change in Coverage sub-index percentile (column 3), change in Resolution sub-index percentile (column 4), and change in Usability sub-index percentile (column 5). Column 1 has no additional controls; columns 2–5 add first-differenced control variables and state fixed effects. The minor sample loss across columns 2–5 reflects missing control values in one of the two endpoint years. Standard errors clustered at the state level appear in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 5: Visibility transformation comparison: raw, z-scored, and within-year percentile rank

	(1) PDV_raw	(2) PDV_z	(3) PDV_pct
PDV_raw	-0.0848		
	(0.0864)		
PDV_z		-0.0042	
		(0.0195)	
PDV_pct			0.0022**
			(0.0010)
Obs	34162	34162	34162
Within R-sq	0.002	0.001	0.002

Notes. The headline regression re-estimated with three alternative transformations of the Public Data Visibility composite to test the model's rank-versus-level prediction (Proposition 3). Column 1 uses the raw composite `PDV_raw` on its 0–4 ordinal scale. Column 2 uses the within-year z-score `PDV_z`. Column 3 uses the within-year percentile rank `PDV_pct` on a 0–100 scale. The dependent variable in all three columns is the inverse hyperbolic sine of real-2022-dollar per-capita competitive grants. All specifications include county and state-by-year fixed effects plus the full control vector. Standard errors clustered at the county level appear in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 6: Sub-index decomposition of the visibility effect on competitive grant capture

	(1) Coverage only	(2) Resolution only	(3) Usability only	(4) All three
PDV_coverage_pct	0.0022***			0.0027***
	(0.0008)			(0.0010)
PDV_resolution_pct		0.0003		0.0011
		(0.0007)		(0.0009)
PDV_usability_pct			-0.0002	-0.0003
			(0.0008)	(0.0008)
Observations	34170	34170	34170	34170
Within R-sq	0.003	0.002	0.003	0.003

Notes. Each column reports a separate two-way fixed-effects regression. The dependent variable is the inverse hyperbolic sine of real-2022-dollar per-capita competitive grants. The regressors are the within-year percentile ranks of the three PDV sub-indices: Coverage (the share of the eight domains in which the county has any measurable data, capturing the extensive margin of visibility), Resolution (the spatial granularity of available data), and Usability (the technical accessibility of the data). Columns 1–3 enter each sub-index alone; column 4 enters all three jointly. All specifications include county and state-by-year fixed effects plus the full control vector. Standard errors clustered at the county level appear in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 7: Domain-level visibility, single-domain and joint specifications

	(1) Health	(2) Env	(3) Broadband	(4) Housing	(5) Transp	(6) Schools	(7) Finance	(8) Bus/Lab	(9) joint
score_health_pct	0.0025***								0.0023***
	(0.0008)								(0.0008)
score_environment_pct		0.0029***							0.0023**
		(0.0009)							(0.0009)
score_broadband_pct			0.0000						0.0000
			(.)						(.)
score_housing_pct				0.0025					0.0025
				(0.0028)					(0.0028)
score_transportation_pct					0.0004				0.0005
					(0.0008)				(0.0008)
score_schools_pct						0.0062***			0.0061***
						(0.0009)			(0.0009)
score_local_finance_pct							-0.0005		-0.0004
							(0.0007)		(0.0007)
score_business_labor_pct								0.0000	0.0000
								(.)	(.)
Obs	34170	34170	34170	34170	34170	34170	34170	34170	34170
Within R-sq	0.002	0.002	0.002	0.002	0.002	0.002	0.004	0.002	0.004

Notes. Each of columns 1–8 reports a separate two-way fixed-effects regression of the inverse hyperbolic sine of real-2022-dollar per-capita competitive grants on a single domain percentile rank; column 9 enters all eight domain percentile ranks jointly. The eight domains are the substantive policy areas underlying the PDV composite. The Broadband and Business/Labor coefficients are reported as 0.0000 with "(.)" standard errors because the underlying domain scores have insufficient within-year cross-county variation for identification: FCC broadband reporting is uniform across counties in any given year, and the CBP-based business/labor score is time-stable. All specifications include county and state-by-year fixed effects and the full control vector. Standard errors clustered at the county level appear in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 8: Instrumental variables estimates: state open-data policies and the Bartik shift-share predictor

	(1) FS: PDV_pct	(2) IV 1-instr	(3) FS: PDV_pct	(4) IV 2-instr	(5) IV 2 st-trends	(6) FS: Bartik IV	(7) IV: Bartik
state_od_law	2.0028***		2.2172***				
	(0.4234)		(0.4372)				
state_cdo			1.5199***				
			(0.4280)				
PDV_bartik						38.4272***	
						(5.2167)	
log_pop	7.4521**	-0.6510	7.0814*	-0.6663*	-0.8215	7.4708**	0.3537
	(3.6762)	(0.4144)	(3.6604)	(0.4017)	(0.6760)	(3.2599)	(0.4142)
unemp_rate	-0.0270	-0.0365**	-0.0599	-0.0364**	-0.0650***	0.4128***	0.0035
	(0.1263)	(0.0146)	(0.1261)	(0.0147)	(0.0175)	(0.1069)	(0.0172)
poverty_pct	-0.2005***	-0.0038	-0.2050***	-0.0033	-0.0283**	-0.1316**	-0.0101
	(0.0650)	(0.0091)	(0.0648)	(0.0086)	(0.0135)	(0.0552)	(0.0086)
share_black	11.7737	-0.0130	12.0828	-0.0422	-1.3867	-0.7398	-1.3684
	(24.0398)	(2.4018)	(23.9670)	(2.4007)	(4.8466)	(21.3223)	(2.8355)
share_hisp	-8.0574	1.8452	-8.7673	1.8639	-1.5321	4.7631	0.8119
	(15.4291)	(2.3052)	(15.3987)	(2.3029)	(3.4774)	(12.9946)	(2.1393)
share_65plus	199.0489***	3.0638	194.9520***	2.6095	14.8873***	144.7411***	18.5399***
	(22.6457)	(4.9731)	(22.3899)	(4.0511)	(4.5636)	(18.1604)	(3.6357)
median_inc_10k	0.1498	-0.0567	0.1674	-0.0571	-0.0653	0.2026	-0.0330
	(0.3149)	(0.0349)	(0.3127)	(0.0348)	(0.0619)	(0.2599)	(0.0370)
dem_voteshare_pres	-7.9891***	0.2398	-8.1500***	0.2583	-1.5593***	-3.8634**	-0.2422
	(1.7835)	(0.2486)	(1.7777)	(0.2224)	(0.5446)	(1.5558)	(0.2202)
n_disasters	0.0302	0.0123	0.0958	0.0122	0.0851***	0.0851	0.0127
	(0.1168)	(0.0155)	(0.1185)	(0.0155)	(0.0219)	(0.0811)	(0.0130)
PDV_pct		0.0267**		0.0290**	0.1855***		0.0775***
		(0.0132)		(0.0140)	(0.0443)		(0.0210)
Observations	23636	23636	23636	23636	23636	34170	34170
Weak-IV F		22.37		15.13	14.97		54.26

Notes. Two-stage least squares estimates of the causal effect of public data visibility on competitive federal grant capture. Columns 1, 3, and 6 report first-stage regressions of `PDV_pct` on the instruments; columns 2, 4, 5, and 7 report second-stage regressions in which the dependent variable is the inverse hyperbolic sine of real-2022-dollar per-capita competitive grants and the instrumented regressor is `PDV_pct` (its predicted value from the relevant first stage). Column 2 uses a single instrument (`state_od_law`, an indicator for state-level open-data statute or executive order in effect in year t); columns 4 and 5 add `state_cdo`, an indicator for a state Chief Data Officer position; column 7 uses the Bartik shift-share predictor `PDV_bartik` constructed from 2012 county-level baseline exposure to each PDV domain interacted with national domain trends. Columns 1–5 are restricted to states with high- or medium-confidence open-data dating (23,636 county-year observations); columns 6–7 use the full panel (34,170 observations). All specifications include county and year fixed effects, with the addition of state-specific linear time trends in column 5. The Weak-IV F row reports the Kleibergen–Paap robust first-stage F-statistic. Standard errors clustered at the county level appear in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 9: Placebo battery: visibility rank across federal grant categories

	(1) disc_clean	(2) Broad disc	(3) Formula	(4) Direct	(5) Loans	(6) Total
PDV_pct	0.0022**	0.0006	0.0005	-0.0001	0.0009	-0.0000
	(0.0010)	(0.0011)	(0.0015)	(0.0001)	(0.0008)	(0.0002)
Obs	34162	34162	34162	34162	34162	34162
Within R-sq	0.002	0.002	0.000	0.024	0.011	0.007

Notes. The headline regression re-estimated across six federal grant outcomes to test the model's mechanism-specificity prediction (Proposition 4). The dependent variable in each column is the inverse hyperbolic sine of real-2022-dollar per-capita obligations from the indicated grant category: `disc_clean` (column 1) is the 34-program curated competitive panel that serves as the headline outcome; `Broad disc` (column 2) is the full set of USAspending `assistance_type_code` 04 (Project Grant) + 05 (Cooperative Agreement) awards, which includes quasi-formula programs excluded from the curated panel; `Formula` (column 3) is codes 02 (Block Grant) + 03 (Formula Grant); `Direct` (column 4) is codes 06–07 (direct payments to individuals); `Loans` (column 5) is code 10 (direct loans); `Total` (column 6) is the sum of all federal financial assistance excluding insurance and guaranteed loans. All specifications include county and state-by-year fixed effects plus the full control vector, and the regressor of interest is the within-year PDV percentile rank. Standard errors clustered at the county level appear in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 10: Heterogeneity by county population quartile

	(1) Q1 smallest	(2) Q2	(3) Q3	(4) Q4 largest
PDV_pct	-0.0005	0.0040**	0.0030	0.0010
	(0.0016)	(0.0019)	(0.0019)	(0.0019)
Obs	8425	8580	8600	8536
Within R-sq	0.005	0.003	0.004	0.005

Notes. The headline two-way fixed-effects specification re-estimated separately for four county-size subsamples. Population quartiles are computed within each fiscal year using Census Population Estimates Program annual county population, with cutoffs at the 25th, 50th, and 75th within-year percentiles; counties may shift quartiles across years as their relative size changes. The dependent variable is the inverse hyperbolic sine of real-2022-dollar per-capita competitive grants. Each specification includes county and state-by-year fixed effects plus the full control vector. Standard errors clustered at the county level appear in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 11: Heterogeneity by policy era (DATA Act and COVID-19)

	(1) Pre-DATA	(2) Post-DATA	(3) Pre-COVID	(4) Post-COVID
PDV_pct	0.0028**	0.0002	0.0036***	0.0023
	(0.0014)	(0.0014)	(0.0012)	(0.0026)
Obs	15525	18636	24844	9318
Within R-sq	0.005	0.001	0.004	0.001
Group Equality		0.018		0.046

Notes. The headline two-way fixed-effects specification re-estimated separately for four era subsamples. The Digital Accountability and Transparency Act's full compliance deadline of May 2017 defines the pre-DATA-Act sample (fiscal years 2012–2016, column 1) and the post-DATA-Act sample (fiscal years 2017–2022, column 2). The COVID-19 split is at fiscal year 2020: pre-COVID is 2012–2019 (column 3) and post-COVID is 2020–2022 (column 4). The dependent variable is the inverse hyperbolic sine of real-2022-dollar per-capita competitive grants. Each specification includes county and state-by-year fixed effects plus the full control vector. Standard errors clustered at the county level appear in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 12: Heterogeneity by county racial and ethnic composition

	(1) Low Black	(2) High Black	(3) Low Hisp	(4) High Hisp	(5) Low AIAN	(6) High AIAN
PDV_pct	0.0022**	0.0025	0.0019*	0.0038	0.0021**	0.0042
	(0.0011)	(0.0025)	(0.0011)	(0.0030)	(0.0010)	(0.0046)
Obs	27245	6874	28369	5704	31917	2164
Within R-sq	0.002	0.002	0.002	0.003	0.002	0.008
Group Equality		0.902		0.568		0.641

Notes. The headline two-way fixed-effects specification re-estimated separately for six demographic subsamples. Counties are classified as "high" if the Census Population Estimates Program share of the relevant population group exceeds the indicated cutoff in any panel year: high-Black if `share_black` > 0.15, high-Hispanic if `share_hisp` > 0.15, and high-AIAN (American Indian / Alaska Native) if `share_aian` > 0.05. The corresponding "low" subsamples are the complement of each high subsample (counties at or below the cutoff). The high-AIAN cutoff is set lower than the others because AIAN shares are sharply concentrated in tribal-land counties; the resulting high-AIAN subsample contains 2,164 county-year observations. The dependent variable is the inverse hyperbolic sine of real-2022-dollar per-capita competitive grants. Each specification includes county and state-by-year fixed effects plus the full control vector. Standard errors clustered at the county level appear in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 13: Robustness: alternative outcome transformations

	(1) IHS pc	(2) log(1+pc)	(3) Level pc	(4) Any (1/0)	(5) Share	(6) log alloc	(7) N progs
PDV_pct	0.0022**	0.0021**	0.0479**	0.0006***	0.0000*	0.0018*	0.0018***
	(0.0010)	(0.0009)	(0.0195)	(0.0002)	(0.0000)	(0.0011)	(0.0006)
Obs	34162	33985	34170	34162	34162	14468	34162
Within R-sq	0.002	0.002	0.001	0.002	0.002	0.005	0.007

Notes. The headline regression re-estimated across seven alternative transformations of the competitive-grants outcome, all regressed on the within-year PDV percentile rank. Column 1 (IHS pc) is the headline specification: inverse hyperbolic sine of real-2022-dollar per-capita competitive grants. Column 2 replaces IHS with $\log(1 + \widetilde{G}^r)$, dropping county-years with negative real-dollar values (refunds and clawbacks). Column 3 uses the untransformed level (real-2022 dollars per capita). Column 4 is the extensive-margin indicator (1 if the county received any competitive grant in the fiscal year). Column 5 is the county's within-year share of national competitive-grant dollars. Column 6 is the natural log of the county's allocation ratio (grant share / population share), defined only for county-years with positive shares (hence the reduced sample). Column 7 is the count of distinct programs (out of 34) from which the county received any obligation. All specifications include county and state-by-year fixed effects plus the full control vector. Standard errors clustered at the county level appear in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 14: Robustness: subsample restrictions

	(1) Full	(2) No tribal	(3) No top metro
PDV_pct	0.0022**	0.0021**	0.0025**
	(0.0010)	(0.0010)	(0.0011)
Obs	34162	32928	25605
Within R-sq	0.002	0.002	0.001

Notes. The headline two-way fixed-effects specification re-estimated on three subsamples. Column 1 (Full) reports the headline result on the full panel for comparison. Column 2 (No tribal) drops counties whose Census Population Estimates Program AIAN share exceeds 10% in any panel year, removing 1,234 county-year observations. Column 3 (No top metro) drops counties in the top within-year population quartile, restricting attention to roughly the bottom 75% of the county size distribution and removing 8,557 county-year observations. The dependent variable is the inverse hyperbolic sine of real-2022-dollar per-capita competitive grants. All specifications include county and state-by-year fixed effects plus the full control vector. Standard errors clustered at the county level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 15: Variable Definitions

Variable	Label	Definition
Panel A: Public Data Visibility (PDV) Index		
`PDV raw`	PDV Composite (raw)	County-year composite Public Data Visibility score on a 0–4 ordinal scale; simple average of eight domain scores (Health, Environment, Broadband, Housing, Transportation, Schools, Local Finance, Business/Labor).
`PDV z`	PDV Composite (z-score)	Within-year standardized PDV composite; mean zero and unit standard deviation within each fiscal year.
`PDV pct`	PDV Percentile Rank	Within-year percentile rank of the PDV composite (0–100); the headline regressor.
`PDV coverage pct`	Coverage Sub-index Rank	Within-year percentile rank of the Coverage sub-index: share of the eight domains for which the county has any publicly available federal data (extensive margin of visibility).
`PDV resolution pct`	Resolution Sub-index Rank	Within-year percentile rank of the Resolution sub-index: spatial granularity of available data.
`PDV usability pct`	Usability Sub-index Rank	Within-year percentile rank of the Usability sub-index: technical accessibility of the data.
`score health pct`	Health Domain Rank	Within-year percentile rank of the Health domain score.
`score environment pct`	Environment Domain Rank	Within-year percentile rank of the Environment domain score.
`score broadband pct`	Broadband Domain Rank	Within-year percentile rank of the Broadband domain score. Near-zero within-year variation due to uniform FCC broadband reporting across counties.

Continued on next page

Table 15 continued

Variable	Label	Definition
score housing pct	Housing Domain Rank	Within-year percentile rank of the Housing domain score.
score transportation pct	Transportation Domain Rank	Within-year percentile rank of the Transportation domain score.
score schools pct	Schools Domain Rank	Within-year percentile rank of the Schools domain score.
score local finance pct	Local Finance Domain Rank	Within-year percentile rank of the Local Finance domain score.
score business labor pct	Business/Labor Domain Rank	Within-year percentile rank of the Business/Labor domain score. Near-zero within-year variation due to time-stability of CBP-based scores.
PDV pct l1	PDV Rank (lagged)	One-year lag of PDV pct.
PDV bartik	PDV Bartik Predictor	Bartik shift-share instrument: 2012 county-level baseline exposure to each PDV domain interacted with national domain trends; used as an IV for PDV pct.
Panel B: Competitive Federal Grant Outcomes		
grants disc clean pc real22	Competitive Grants (IHS)	Inverse hyperbolic sine of real 2022-dollar per-capita competitive federal grant obligations from 34 hand-curated programs across eleven agencies; the headline dependent variable.
n programs disc clean	Program Count	Count of distinct programs (out of 34) from which the county received any obligation in the fiscal year.
any disc clean	Any Competitive Grant	Indicator equal to one if the county received any competitive grant obligation in the fiscal year.

Continued on next page

Table 15 continued

Variable	Label	Definition
Panel C: Placebo Grant Categories		
`disc clean`	Curated Competitive	IHS of per-capita obligations from the 34-program curated panel (headline outcome).
`Broad disc`	Broad Discretionary	IHS of per-capita obligations from USASpending type codes 04 (Project Grant) and 05 (Cooperative Agreement), including quasi-formula programs excluded from the curated panel.
`Formula`	Formula Grants	IHS of per-capita obligations from type codes 02 (Block Grant) and 03 (Formula Grant).
`Direct`	Direct Payments	IHS of per-capita obligations from type codes 06–07 (direct payments to individuals).
`Loans`	Direct Loans	IHS of per-capita obligations from type code 10 (direct loans).
`Total`	Total Federal Assistance	IHS of per-capita total federal financial assistance, excluding insurance and guaranteed loans.
Panel D: Control Variables		
`log pop`	Log Population	Natural log of county annual population from Census Population Estimates Program (PEP).
`unemp rate`	Unemployment Rate	County annual unemployment rate (percent) from BLS Local Area Unemployment Statistics (LAUS).
`poverty pct`	Poverty Rate	County share of population below the federal poverty line (percent) from Census SAIPE.
`median inc 10k`	Median Household Income	County median household income in real 2022 dollars scaled to $10,000 units, from Census SAIPE.

Continued on next page

Table 15 continued

Variable	Label	Definition
`share black`	Black Population Share	County share of population identifying as Black or African American alone, from Census PEP.
`share hisp`	Hispanic Population Share	County share of population identifying as Hispanic or Latino, from Census PEP.
`share 65plus`	Elderly Population Share	County share of population aged 65 and older, from Census PEP.
`dem voteshare pres`	Dem. Presidential Vote Share	County two-party Democratic presidential vote share from MIT Election Data and Science Lab (MEDSL).
`n disasters`	FEMA Disaster Declarations	Count of FEMA major disaster declarations active in the county-year from FEMA OpenFEMA.
Panel E: Instrumental Variables		
`state od law`	State Open-Data Law	Indicator equal to one if a state-level open-data statute or executive order was in effect in year t.
`state cdo`	State Chief Data Officer	Indicator equal to one if the state had an active Chief Data Officer position in year t.
Panel F: Long-Difference Variables		
`d PDV pct`	Δ PDV Rank	2022-minus-2012 change in the within-year PDV percentile rank.
`d PDV coverage pct`	Δ Coverage Rank	2022-minus-2012 change in the within-year Coverage sub-index percentile rank.
`d PDV resolution pct`	Δ Resolution Rank	2022-minus-2012 change in the within-year Resolution sub-index percentile rank.

Continued on next page

Table 15 continued

Variable	Label	Definition
`d PDV usability pct`	Δ Usability Rank	2022-minus-2012 change in the within-year Usability sub-index percentile rank.

Notes. The analysis panel covers 3,144 U.S. counties and county-equivalents over fiscal years 2012–2022 (34,533 county-year observations). PDV variables are constructed for this paper; all other variables are drawn from publicly available federal and administrative sources. County identifiers are 5-digit zero-padded FIPS codes. Real-dollar variables are deflated to 2022 using the BLS CPI-U.

Technical Appendix A:

Data Construction for the Public Data Visibility (PDV) Index

Prepared for:

"The Legibility Premium: Public Data Visibility and the Allocation of Competitive Federal Grants"

May 2026

1. Overview

This appendix documents the complete construction of the Public Data Visibility (PDV) index, a county-by-year measure of the quality and accessibility of publicly available administrative data across eight substantive domains. The index quantifies how well a federal statistical agency, researcher, or policymaker could characterize local conditions in county i during calendar year t using publicly available data alone.

The construction pipeline follows a deliberate three-tier architecture: a *source registry* that formally catalogs all planned data sources; *domain evidence scripts* that translate source availability into standardized evidence rows; and a *scoring and aggregation layer* that applies a rubric to produce the final composite index. The architectural separation ensures that source-level decisions (what data exist, for which counties, in what format) are fully documented before any scoring takes place, and that the rubric is applied uniformly across all domains.

The index covers **8 core domains** ($D = 8$): Health, Environment, Broadband, Housing, Transportation, Schools, Local Finance, and Business/Labor. Five domains employ *AND-logic* overrides—a requirement that data from *both* a primary and a complementary secondary source exist before a county earns the highest score $S = 4$; meeting only the primary criterion caps that domain at $S = 3$. This prevents a single universally-available source from eliminating cross-county variation and ensures that the top score reflects a genuinely richer data environment. The five AND-logic domains are Environment, Broadband, Housing, Schools, and Local Finance. The Transportation domain instead uses a three-tier scoring structure based on two independent sources (NHTSA FARS and FTA NTD).

The final dataset (`pdv_county_year.dta`) contains 34,574 observations spanning 3,144 counties and 11 calendar years (2012–2022). Section 2 describes the source registry; Section 3 details the three-tier workflow; Sections 4–7 document the panel, rubric, and scoring formulas; Section 8 details each domain; Section 9 characterizes cross-sectional variation; Section 10 provides the variable codebook; and Section 11 presents summary statistics. All tables are collected at the end of this appendix.

2. Source Registry

2.1. Purpose and Design

The source registry (`config/sources_registry.csv`) is the authoritative planning document for the PDV project. It catalogs every data source that the project has identified as relevant—whether currently implemented or used only as a support input—and records a standardized

set of metadata about each source's accessibility, format, geographic grain, and role in the PDV scoring framework.

The registry serves three concrete functions. First, it forces explicit decisions about source scope *before* writing any evidence code: each row must answer whether data are publicly accessible, whether a documented API exists, what the primary geographic grain is, and what download strategy is appropriate. Second, it defines the contract between the planning layer and the evidence layer: each row's `source_id` field is used as the filename stem for the corresponding evidence CSV (`data/interim/{source_id}_evidence.csv`), ensuring that every script's output is traceable back to a registry entry. Third, it distinguishes core domain sources (which receive a PDV score) from support sources (which feed the panel shell, control variables, or outcome variables but are not scored).

2.2. Registry Fields

Table 1 describes each column in `sources_registry.csv`.

2.3. Full Source Catalog

The registry contains 16 core domain sources spanning all 8 PDV domains (two sources per domain for Environment, Broadband, Housing, Transportation, Schools, Local Finance, and Business/Labor; one source for Health). Table 2 presents the complete catalog.

2.4. Key Design Decisions

Three important design decisions structure the source selection and scoring methodology:

- **AND-logic for five domains.** For Environment, Broadband, Housing, Schools, and Local Finance, the pipeline introduces a mandatory two-source requirement: a county must have data from a primary *and* a complementary secondary source to qualify for $S = 4$. Primary-only counties are capped at $S = 3$. This prevents universally available sources from eliminating cross-sectional variation and ensures that the highest score reflects a richer, more complete data environment (see Section 5.5 for domain-specific rules).
- **Three-tier transportation scoring.** The transportation domain uses two independent sources: FTA NTD (transit-agency coverage) and NHTSA FARS (fatal crash records). Via MAX aggregation, the resulting distribution is $\{0, 3, 4\}$: counties with no fatal crashes and no transit agency score $S = 0$; counties with FARS crash records but no NTD agency score $S = 3$; counties with NTD transit agencies score $S = 4$.

- **Broadband: FCC sources supplemented by ACS subscriptions.** FCC Form 477 provides block-level deployment data for all counties from 2014 onward, but lacks a documented county-query REST API through 2021. ACS Table B28002 (internet subscriptions at county level) provides a complementary access-demand signal with a documented Census API. The AND-logic constraint requires both FCC deployment data and ACS subscription data to qualify for $S = 4$ in FCC-only filing years (2014–2019).

3. Overall Workflow and Three-Tier Architecture

3.1. Design Philosophy

The PDV construction pipeline is organized around a strict separation of concerns across three tiers. The first tier is the *planning layer* (the source registry), which documents what data sources exist, in what formats, at what geographic grain, and with what access requirements — before any code is written. The second tier is the *evidence construction layer*, which translates source availability into a uniform evidence representation. The third tier is the *scoring and aggregation layer*, which applies the rubric and computes the composite index.

This three-tier separation has four methodological advantages. First, it makes source-level decisions auditable and reproducible: all judgments about whether a source has a documented API, what its geographic grain is, and what its freshness lag is are recorded in the evidence CSVs, not embedded in scoring code. Second, it allows the scoring rubric to be changed without re-running any data collection: re-running only `11_score_domains.py` and downstream scripts is sufficient to propagate a rubric change across all domains. Third, it supports the paper's use of the index as a quasi-objective institutional measure: the registry entries and evidence notes provide documentary proof that scoring decisions were made based on source characteristics, not on the outcome being studied. Fourth, it makes the pipeline extensible: adding a new source requires only writing a new evidence script that conforms to the 16-column `EVIDENCE_COLS` schema, then re-running the scoring layer.

3.2. Workflow Diagram

Figure 1 illustrates the full data flow from the source registry through to the final Stata dataset.

3.3. Tier 1: Source Registry (Planning Layer)

The source registry is written and audited *before* evidence scripts are written. For each source, the analyst must answer the following questions and record the answers in the registry:

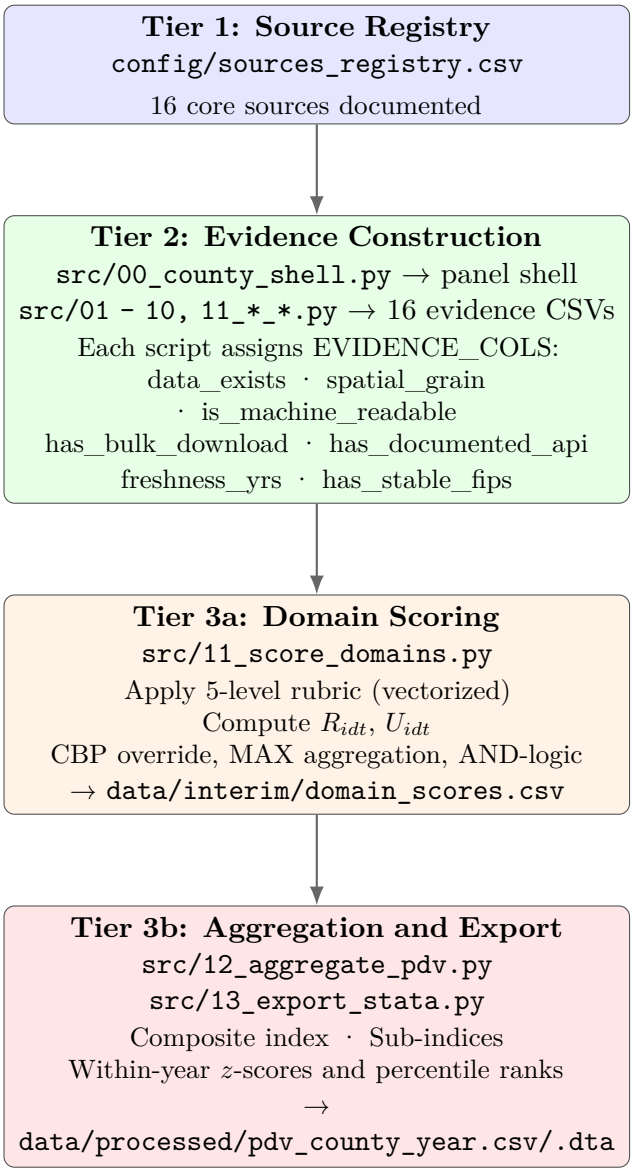

Figure 1: PDV data pipeline: three-tier architecture from source registry to final Stata dataset.

1. **Access**: Is the source publicly accessible? Is a key or account required? Can bulk downloads bypass registration?
2. **Format**: Is the source machine-readable (CSV, JSON, shapefile)? Is there a documented REST API or FTP endpoint?
3. **Grain**: What is the primary geographic grain? Is there a sub-county grain available? Do records contain stable FIPS or GEOID identifiers that allow county-level aggregation?
4. **Freshness**: How frequently is the source updated? What is the typical lag between the reference data year and public release?
5. **Years**: For which panel years (2012–2022) is data expected to exist?
6. **Strategy**: What is the operational download or API strategy? Are there known limitations, archival gaps, or versioning issues?

3.4. Tier 2: Evidence Construction Layer

The evidence construction layer consists of 16 Python scripts that translate raw source data into standardized evidence rows. Each script follows an identical structure: (1) load the county×year panel; (2) load pre-built lookup files or download source data directly; (3) for each panel row, assign all 16 `EVIDENCE_COLS` fields based on source characteristics; (4) save the evidence CSV to `data/interim/{source_id}_evidence.csv`.

A critical design principle is that evidence scripts do *not* score the data. They record factual characteristics of the source (does data exist for this county in this year? what is the geographic grain? is there a documented API?) but do not apply the rubric. Scoring is entirely deferred to Tier 3.

Interim lookup files. Several evidence scripts require county-level information pre-built from external downloads:

- `ntd_county_coverage.json`: maps each panel year to the set of county FIPS codes with at least one NTD-reporting transit agency, constructed via a ZIP-to-ZCTA-to-county crosswalk applied to NTD agency ZIP codes.
- `tri_county_fips.json`: the set of county FIPS codes with at least one TRI-reporting industrial facility, queried from the EPA EFservice API.
- `aqs_county_year.csv`: county-year pairs with at least one EPA AQS air quality monitor, built from annual bulk CSV downloads at https://aqs.epa.gov.
- `fhfa_county_hpi.csv`: county-year pairs with a valid FHFA All-Transactions House Price Index value, built from the FHFA experimental county HPI file.
- `fars_county_year.csv`: county-year pairs with at least one NHTSA fatal crash record,

built from annual NHTSA FARS national CSV ZIP files.

- `f33_county_completeness.csv`: county-year mean district reporting fraction from the NCES F-33 District Finance Survey, constructed by matching district CONUM codes to county FIPS and computing the share of revenue line items reported (flag="R") per county-year.
- `cbp_estab_2019.json`: maps county FIPS to total establishment count (NAICS "00") from the 2019 Census CBP API.

3.5. Tier 3: Scoring and Aggregation Layer

The scoring and aggregation layer consists of three scripts that operate entirely on the evidence CSV files produced in Tier 2.

Script `11_score_domains.py`: Domain Scoring.

1. **Load all 16 evidence files.** Read all evidence CSVs and coerce boolean columns (`data_exists`, `county_specific`, `is_machine_readable`, `has_documented_api`, `has_bulk_download`, `has_stable_fips`) to proper Python booleans.
2. **Apply the general rubric (vectorized).** The function `score_dataframe()` applies the five-level scoring cascade (Section 5) using vectorized pandas boolean operations across all rows simultaneously.
3. **Compute R_{idt} and U_{idt}.** Resolution and usability sub-scores are computed via vectorized operations.
4. **Apply the CBP domain override.** Counties with total CBP establishment count $\geq$ 2,000 (proxy for low suppression rate) are upgraded to $S = 4$ in the business/labor domain.
5. **Aggregate to best score per cell (MAX).** Where multiple sources cover the same domain, take the maximum S_{idt}, R_{idt}, and U_{idt} across sources for each county×year×domain cell.
6. **Apply AND-logic overrides** (Section 5.5). After MAX aggregation, cap Environment, Broadband, Housing, Schools, and Local Finance at $S = 3$ for counties that qualify on only the primary source without the required secondary source.
7. **Output.** Write `domain_scores.csv` with one row per county×year×domain cell.

Script `12_aggregate_pdv.py`: Aggregation. Pivots the long-format domain scores to wide format, fills missing domain-year combinations with 0, merges county metadata, computes the composite index and sub-indices (Section 7), and computes within-year z-scores and percentile ranks for the composite, all sub-indices, and all 8 domain scores.

4. Panel Structure

4.1. County Universe

The county roster is drawn from the 2020 American Community Survey (ACS) 5-year estimates via the Census Bureau API (https://api.census.gov/data/2020/acs/acs5), restricting to the 50 states and the District of Columbia. Independent cities, census areas, and boroughs that do not appear consistently across federal administrative datasets are excluded. The resulting universe contains **3,144 unique counties**. Population estimates (`pop_2020`) are ACS 2020 5-year county-level totals.

4.2. Panel Dimensions

- Years: 2012–2022 (11 calendar years)
- Counties: 3,144
- Total observations: $3{,}144 \times 11 = 34{,}574$ county×year rows

County FIPS codes (`county_fips`) are zero-padded five-digit strings. State FIPS codes (`state_fips`) are zero-padded two-digit strings. Both identifiers are stable throughout the panel. Script: `src/00_county_shell.py`.

5. PDV Scoring Rubric

For each county i, domain d, and year t, we assign a domain score $S_{idt} \in \{0, 1, 2, 3, 4\}$ according to the ordinal rubric in Table 3.

5.1. Evidence Fields

Each evidence row encodes eight binary or continuous fields that map directly onto the rubric:

Field	Type	Description
`data_exists`	bool	Any public data for domain d in year t
`county_specific`	bool	Data specific to county i (not statewide)
`spatial_grain`	int	0=none, 1=state, 2=county, 3=tract/ZIP, 4=block group+
`is_machine_readable`	bool	CSV/JSON/shapefile accessible
`has_bulk_download`	bool	Bulk file at a stable URL
`has_documented_api`	bool	REST API or FTP with documented stable endpoint
`freshness_yrs`	float	Years elapsed since most recent underlying data
`has_stable_fips`	bool	County FIPS or GEOID present in all records

5.2. Scoring Cascade

The general scoring cascade, applied in priority order (highest first), maps evidence fields to the five score levels. The conditions for each level are applied exclusively (first matching rule wins):

$$S_{idt} = 4 \quad \text{if} \quad \texttt{county\_specific} \wedge \texttt{spatial\_grain} \geq 3 \wedge \texttt{has\_documented\_api} \wedge \texttt{freshness} \leq 2 \wedge \texttt{has\_stable\_fips} \tag{1}$$

$$S_{idt} = 3 \quad \text{if} \quad \texttt{county\_specific} \wedge \texttt{is\_machine\_readable} \wedge \texttt{has\_bulk\_download} \wedge \texttt{freshness} \leq 5 \tag{2}$$

$$S_{idt} = 2 \quad \text{if} \quad \texttt{county\_specific} \wedge \Big(\texttt{freshness} > 5 \vee \neg\texttt{is\_machine\_readable} \vee \neg\texttt{has\_bulk\_download}\Big) \tag{3}$$

$$S_{idt} = 1 \quad \text{if} \quad \texttt{data\_exists} \wedge \neg\texttt{county\_specific} \tag{4}$$

$$S_{idt} = 0 \quad \text{if} \quad \neg\texttt{data\_exists} \tag{5}$$

5.3. MAX Aggregation Across Sources Within Domain

Where multiple sources cover the same domain, the pipeline takes the maximum S_{idt}, R_{idt}, and U_{idt} across all sources for each county×year×domain cell. This means a county benefits from its best available data source. For example, in the Transportation domain, a county with both FARS crash records and an NTD transit agency would take $\max(S_{\text{FARS}}, S_{\text{NTD}}) = \max(3, 4) = 4$.

5.4. Domain-Specific $S = 4$ Override: Business/Labor

Counties where the Census County Business Patterns suppression rate is below 30% of four-digit NAICS cells receive $S = 4$. This is proxied by total establishment count: $\texttt{ESTAB} \geq 2{,}000$ implies low suppression. Establishment counts are drawn from the 2019 CBP API and applied as a time-stable proxy across all panel years.

5.5. AND-Logic Overrides

For five domains, the general rubric would assign $S = 4$ to all (or nearly all) counties based on a single universally-available source, eliminating cross-sectional variation. To preserve

meaningful within-year variation while retaining theoretical validity, an AND-logic constraint is applied *after* MAX aggregation: a county must qualify on both a primary source *and* a secondary source to retain $S = 4$. Counties that qualify only on the primary source are capped at $S = 3$.

5.5.1. Environment: EJScreen Alone → Capped at $S = 3$

EPA EJScreen provides block-group-level modelled environmental indicators for *all* counties from 2015 onward, yielding $S = 4$ universally under the general rubric. However, EJScreen scores are derived estimates, not direct measurements. Full environmental data legibility requires corroboration from at least one source of directly measured or reported data.

Rule: A county retains $S = 4$ in the environment domain if and only if it has EJScreen *and* at least one of: (a) TRI facility records (`environment_tri`) or (b) EPA AQS air quality monitor measurements (`environment_aqs`). Counties with EJScreen only are capped at $S = 3$.

Result: 3,354 county-years are capped at $S = 3$ (EJScreen only; rural counties with no TRI facilities and no AQS monitors); 29,977 county-years retain $S = 4$; 1,243 county-years score $S = 0$ (no EJScreen pre-2015 and no TRI, concentrated in 2012–2014).

5.5.2. Broadband: FCC Alone → Capped at $S = 3$

FCC Form 477 provides Census block-level broadband availability data for all U.S. counties from 2014 onward. However, FCC deployment data do not have a documented county-query REST API in the Form 477 era (2014–2021), so they satisfy only the $S = 3$ bulk-download criterion. The ACS Table B28002 (county-level internet subscription rates) provides a complementary access-demand measure with a documented Census API.

Rule: In panel years with both FCC block-level data and available ACS B28002 data, a county must have both to qualify for $S = 4$. FCC-only counties (where ACS data are unavailable or incomplete) are capped at $S = 3$. In 2022, FCC Broadband Data Collection (BDC) satisfies all $S = 4$ criteria independently via its documented public API.

Result: $S = 1$: 6,287 cells (2012–2013, state-level FCC only); $S = 3$: 9,443 cells (FCC without ACS supplement, concentrated in 2014–2016); $S = 4$: 18,844 cells (FCC + ACS from 2017 onward; BDC in 2022).

5.5.3. Housing: CHAS Alone → Capped at $S = 3$

HUD CHAS data are available at the tract level for all counties in all panel years, yielding $S = 4$ universally. However, CHAS captures only housing affordability (income relative to housing costs); it contains no housing price or market dynamics information. Full housing data legibility requires a market price index.

Rule: A county retains $S = 4$ in the housing domain if and only if it has CHAS *and* a valid FHFA County All-Transactions House Price Index (`housing_fhfa_hpi`). The FHFA publishes an experimental annual county-level HPI for counties with sufficient mortgage transaction volume. Counties with too few transactions (typically rural counties with thin housing markets, approximately 380–400 per year) do not receive a county-specific HPI and are capped at $S = 3$.

Result: 4,244 county-years are capped at $S = 3$ (CHAS only, no FHFA county HPI); 30,330 county-years retain $S = 4$ (CHAS + FHFA HPI).

5.5.4. Schools: CCD Alone → Capped at $S = 3$

NCES Common Core of Data (CCD) provides school directory and enrollment records for every public school district in every county in all panel years, yielding $S = 4$ universally. However, CCD covers only administrative and enrollment information. Full school data legibility requires that districts also report financial data.

Rule: A county retains $S = 4$ in the schools domain if and only if it has CCD *and* the NCES F-33 District Finance Survey (`schools_nces_f33`) with a mean district reporting fraction ≥ 0.75 (i.e., at least 75% of revenue line items reported across all districts in the county). Counties with CCD only (low F-33 completeness) are capped at $S = 3$.

The F-33 is the annual NCES District Finance Survey that collects per-district revenue, expenditure, and debt data using the standardized Census Bureau financial survey instrument. Each line item carries a flag: "R" (reported) or "M" (missing/not reported). The county-level mean reporting fraction is computed from the district-level flags.

Result: 4,193 county-years are capped at $S = 3$ (CCD only, F-33 completeness below threshold); 30,381 county-years retain $S = 4$ (CCD + high F-33 completeness).

5.5.5. Local Finance: ASGF Alone → Capped at $S = 3$

The Census Annual Survey of State and Local Government Finances (ASGF) provides county-level aggregates for all counties in ASGF survey years, but lacks sub-county geographic

resolution (`spatial_grain` = 2), so it satisfies only the $S = 3$ criterion. The Federal Audit Clearinghouse (FAC) collects Single Audit submissions from government entities expending $\geq$ \$750,000 in federal awards annually under OMB Uniform Guidance §200.501. Counties where at least one government entity files a Single Audit have an additional layer of mandatory, independently audited financial disclosure that corroborates and extends the ASGF survey data.

Rule: In ASGF survey years (all non-CoG years), a county retains $S = 4$ if it has both ASGF data *and* at least one FAC Single Audit filer. ASGF-only counties (no FAC filer in that year) are capped at $S = 3$. In Census of Governments census years (2012, 2017, and 2022), all counties score $S = 4$ regardless of FAC status, because the CoG is a complete enumeration with sub-county geocoding via the Government Master Address File.

FAC data are available from the GSA FAC API for fiscal years 2016 onward (https://api.fac.gov/). For 2013–2015 (pre-FAC API availability), all ASGF-year counties score $S = 3$.

Result: $S = 3$: 10,727 county-years (CoG years: all $S = 4$; ASGF years 2013–2015: all $S = 3$; ASGF years 2016, 2018–2021: counties without FAC filer); $S = 4$: 23,847 county-years (all CoG years; ASGF years with FAC filer from 2016 onward).

6. Resolution and Usability Sub-scores

In addition to S_{idt}, each domain record carries two sub-scores used to compute the resolution and usability sub-indices.

Resolution (R_{idt}). The resolution sub-score equals the `spatial_grain` field when data exist, and zero otherwise:

$$R_{idt} = \texttt{spatial\_grain} \cdot \mathbf{1}[\texttt{data\_exists}] \quad \in \{0, 1, 2, 3, 4\}$$

Levels: 0 = no data, 1 = state, 2 = county, 3 = tract or ZIP code, 4 = block group or finer.

Usability (U_{idt}). The usability sub-score counts how many of four access-format conditions are satisfied:

$$U_{idt} = \mathbf{1}[\texttt{has\_bulk\_download}] + \mathbf{1}[\texttt{has\_documented\_api}] + \mathbf{1}[\texttt{freshness} \leq 2] + \mathbf{1}[\texttt{has\_stable\_fips}] \quad \in$$

When multiple sources cover the same domain, the maximum S_{idt}, R_{idt}, and U_{idt} across sources is taken for each county×year×domain cell before AND-logic is applied.

7. Composite Index and Sub-indices

Composite PDV score.

$$\text{PDV}_{it} = \frac{1}{D}\sum_{d=1}^{D} S_{idt}, \qquad D = 8$$

Stored as `PDV_raw`; range [0, 4].

Within-year standardized forms.

$$\texttt{PDV\_z}_{it} = \frac{\text{PDV}_{it} - \bar{\text{PDV}}_t}{\sigma_t}$$

$$\texttt{PDV\_pct}_{it} = \text{percentile rank of PDV}_{it} \text{ within year } t\text{, scaled 0–100}$$

Within-year z-scores and percentile ranks (`_z` and `_pct` suffixes) are computed for `PDV_raw` and for each sub-index and domain score.

Sub-index: Coverage.

$$\text{Coverage}_{it} = \frac{1}{D}\sum_{d=1}^{D} \mathbf{1}[S_{idt} \geq 2]$$

Sub-index: Resolution.

$$\text{Resolution}_{it} = \frac{1}{D}\sum_{d=1}^{D} R_{idt}$$

Sub-index: Usability.

$$\text{Usability}_{it} = \frac{1}{D}\sum_{d=1}^{D} U_{idt}$$

Table 4 reports `PDV_raw` and the sub-indices by calendar year.

8. Domain Construction

8.1. Health

Primary source: CDC PLACES (formerly 500 Cities).

2012–2015 ($S = 0$). No sub-national behavioral health surveillance data were available

at the county level. The CDC Behavioral Risk Factor Surveillance System (BRFSS) is administered at the state level; county-level small-area estimates were not released for this period.

2016–2019, 500-Cities counties ($S = 4$). The CDC 500 Cities Project, launched December 2016, provided tract-level estimates of 27 chronic disease and health behavior measures for the 500 largest U.S. cities, covering **339 counties** (`data/interim/places_500cities_county_fips.json`). These records satisfy all four $S = 4$ conditions: tract-level grain (`spatial_grain` = 3), documented CDC REST API, freshness $\approx$ 2 years, and stable FIPS identifiers.

2016–2019, remaining counties ($S = 3$). The $3{,}144 - 339 = 2{,}805$ counties outside the 500-Cities footprint have county-level BRFSS small-area estimates without consistent tract-level resolution, satisfying only the $S = 3$ bulk-download condition.

2020–2022, PLACES national launch ($S = 4$). CDC PLACES expanded to nationwide tract-level coverage in December 2020, covering all $\approx$3,143 counties. All counties score $S = 4$ from 2020 onward.

Score distribution: $S = 0$: 12,573 cells (2012–2015); $S = 3$: 11,225 (non-500-Cities counties, 2016–2019); $S = 4$: 10,776 (500-Cities 2016–2019 and all counties 2020–2022).

Sources: CDC PLACES https://www.cdc.gov/places. Script: `src/01_health_cdc_places.py`.

8.2. Environment

Primary sources: EPA EJScreen (block-group level), EPA Toxics Release Inventory (TRI, facility level), and EPA Air Quality System (AQS, monitor point level).

EPA EJScreen. EJScreen provides block-group-level environmental and demographic indicators for all U.S. counties annually since 2015, derived from ACS 5-year estimates. Bulk CSV downloads at a stable EPA FTP URL and an ArcGIS REST API satisfy all four $S = 4$ conditions. Panel years 2012–2014 pre-date EJScreen's launch.

EPA Toxics Release Inventory (TRI). TRI compiles facility-level chemical release data. Only counties with at least one TRI-reporting industrial facility appear in TRI records. Approximately **2,681 counties** have TRI facilities; the remaining $\approx$ 463 counties score $S = 0$ for TRI. TRI data satisfy all four $S = 4$ conditions (facility-point geocodes, documented EPA API, annual release, stable FIPS).

EPA Air Quality System (AQS). AQS provides ground-measured concentrations of criteria pollutants from physical monitors deployed across the country. Approximately **1,044**–

1,094 counties per year host at least one AQS monitor (Table 5). AQS data satisfy all four $S = 4$ conditions (monitor-point geocode with sub-county resolution, documented AQS REST API, annual bulk CSV download, stable FIPS).

AND-logic override. EJScreen alone would yield $S = 4$ for all counties from 2015 onward, eliminating cross-sectional variation. The AND-logic constraint (Section 5.5) caps EJScreen-only counties at $S = 3$: counties must have EJScreen *plus* at least one measured source (TRI or AQS) to retain $S = 4$.

Score distribution (final, post-AND-logic): $S = 0$: 1,243 cells; $S = 3$: 3,354 (EJScreen without TRI or AQS, mainly rural post-2014); $S = 4$: 29,977 (EJScreen + TRI or AQS).

Sources: EPA EJScreen https://www.epa.gov/ejscreen; EPA TRI https://www.epa.gov/toxics-release-inventory-tri-program; EPA AQS https://aqs.epa.gov/aqsweb/documents/data_api.html. Scripts: `src/02_environment_ejscreen.py`, `src/03_environment_tri.py`, `src/11_environment_aqs.py`.

8.3. Broadband

Primary sources: FCC Form 477 (2014–2021), FCC Broadband Data Collection (BDC, 2022), and ACS Table B28002 (internet subscriptions, county level).

2012–2013 ($S = 1$). Before the December 2013 Form 477 filing, FCC broadband deployment statistics were published only at the state level.

2014–2016 ($S = 3$). Form 477 provides Census block-level fixed broadband availability data for all U.S. counties as bulk CSV downloads. No documented REST API for county-level queries existed during this period, preventing $S = 4$ despite sub-county block-level grain. ACS B28002 county subscription data is not yet applied for these early panel years.

2017–2021 ($S = 4$ for most counties). From panel year 2017 onward, ACS Table B28002 provides county-level internet subscription rates with a documented Census REST API, complementing FCC block-level deployment data. The AND-logic constraint requires both FCC deployment data and ACS subscription data: counties with FCC data and ACS B28002 data retain $S = 4$; a small number of counties lacking ACS data coverage (≈ 2/year) remain at $S = 3$.

2022 ($S = 4$). The FCC Broadband Data Collection (BDC), launched for the June 2022 filing, provides address- and location-level availability data with a documented public API, satisfying all four $S = 4$ criteria independently.

Score distribution: $S = 1$: 6,287 cells (2012–2013); $S = 3$: 9,443 cells (2014–2016, all counties; 2017–2022, counties without ACS supplement); $S = 4$: 18,844 cells (2017–2022, FCC + ACS; or FCC BDC in 2022).

Sources: FCC open data https://opendata.fcc.gov; FCC BDC https://broadbandmap.fcc.gov/home; Census ACS https://api.census.gov/data. Scripts: `src/04_broadband_fcc.py`, `src/11_broadband_acs.py`.

8.4. Housing

Primary sources: HUD Comprehensive Housing Affordability Strategy (CHAS) and FHFA County All-Transactions House Price Index (HPI).

HUD CHAS. CHAS data are produced as special tabulations of ACS data by the Census Bureau and published by HUD. They provide tract-level measures of housing cost burden by income group for all counties in all panel years. The HUD User Data API provides documented programmatic access. The data lag from ACS end year to panel year is consistently ≈ 2 years, and tract-level grain satisfies all four $S = 4$ conditions. Table 6 shows the CHAS vintage mapping.

FHFA County House Price Index. The FHFA publishes an experimental All-Transactions HPI at the county level, derived from repeat-sales mortgage transaction data. Counties must have sufficient transaction volume to support a reliable index; approximately 380–400 counties per year (typically rural counties with thin housing markets) do not receive a county-specific HPI value.

AND-logic override. CHAS alone would yield $S = 4$ for all counties in all years, eliminating cross-county variation. The AND-logic constraint caps CHAS-only counties at $S = 3$: full housing data legibility requires both affordability data (CHAS) and a market price index (FHFA HPI).

Score distribution (final, post-AND-logic): $S = 3$: 4,244 county-years (CHAS only; no FHFA county HPI); $S = 4$: 30,330 county-years (CHAS + FHFA HPI).

Sources: HUD CHAS https://www.huduser.gov/portal/datasets/cp.html; FHFA County HPI https://www.fhfa.gov/data/hpi. Scripts: `src/05_housing_hud_chas.py`, `src/11_housing_fhfa.py`.

8.5. Transportation

Primary sources: FTA National Transit Database (NTD) and NHTSA Fatality Analysis Reporting System (FARS).

8.5.1. FTA National Transit Database (NTD)

NTD annual files report service and financial data for all publicly funded transit agencies. County coverage is constructed via a two-step crosswalk: NTD agency ZIP codes are matched to Census ZCTA-to-county relationship files, and the resulting county-to-agency crosswalk is stored in `data/interim/ntd_county_coverage.json`.

Transit counties score $S = 4$ under the general rubric: NTD provides agency-level operating statistics with route-level geocodes at sub-county resolution, bulk Excel/CSV downloads at a stable FTA URL, approximately 1-year data lag, and stable agency and county FIPS identifiers. Table 7 reports NTD county coverage by year.

8.5.2. NHTSA Fatality Analysis Reporting System (FARS)

NHTSA FARS is the mandatory federal census of all fatal motor vehicle crashes in the United States. Every county where at least one road fatality occurred is represented in the annual FARS national CSV files, which are publicly available as bulk downloads from NHTSA. Because FARS is a complete enumeration of fatal crashes (not a sample), coverage is near-universal: approximately 2,840–2,880 of the 3,144 panel counties appear in FARS each year. Counties with no fatal crashes (≈ 263 per year, typically very small or remote counties) score $S = 0$ for FARS.

FARS data are county-level crash counts (`spatial_grain` $= 2$), bulk-downloadable as annual CSV files at a stable NHTSA URL, with an approximately 1-year data lag. Because FARS lacks a documented REST API and is at county (not sub-county) grain, it satisfies the $S = 3$ conditions but not the $S = 4$ conditions under the general rubric.

8.5.3. Three-Tier Scoring via MAX Aggregation

The MAX aggregation across sources produces a natural three-tier transportation score distribution without requiring explicit AND-logic:

- $S = 0$: No FARS record *and* no NTD agency (≈ 263 counties/year; very remote).
- $S = 3$: FARS record but no NTD agency: $\max(3_{\text{FARS}}, 0_{\text{NTD}}) = 3$ ($\approx 1{,}280$ counties/year; rural/suburban).

- $S = 4$: NTD transit agency (with or without FARS): $\max(\cdot, 4_{\text{NTD}}) = 4$ ($\approx$ 1,560 counties/year; urban/transit-served).

Score distribution (pooled, post-MAX): $S = 0$: 2,500 cells; $S = 3$: 18,408 cells; $S = 4$: 13,666 cells.

Sources: FTA NTD https://www.transit.dot.gov/ntd/ntd-data; NHTSA FARS https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/FARS/. Scripts: `src/06_transportation_ntd.py`, `src/11_transportation_fars.py`.

8.6. Schools

Primary sources: NCES Common Core of Data (CCD) and NCES F-33 District Finance Survey.

NCES Common Core of Data (CCD). CCD covers all public schools and school districts annually. Every county has at least one public school district. School-level records contain geocoordinates and county FIPS. All four $S = 4$ conditions are satisfied: school geocodes with county FIPS, NCES EDGE REST API, approximately 1-year annual lag, and bulk ZIP download at a stable NCES URL. CCD scores $S = 4$ for all 3,144 counties in all 11 years.

NCES F-33 District Finance Survey. F-33 is the annual NCES financial survey of all local education agencies (LEAs), collecting per-district revenue, expenditure, and debt data using the standardized Census Bureau financial survey instrument. Each revenue and expenditure line item carries a reporting flag: "R" (data reported) or "M" (missing, not reported).

County-year F-33 completeness is measured as the mean fraction of revenue line items flagged "R" across all districts in the county. Counties where the mean reporting fraction is ≥ 0.75 receive `data_exists=True` in the F-33 evidence; counties below this threshold or without any districts reporting receive `data_exists=False`.

AND-logic override. CCD alone would yield $S = 4$ universally. Full school data legibility requires both administrative data (CCD) and financial data (F-33). Counties with CCD but insufficient F-33 completeness are capped at $S = 3$.

Score distribution (final, post-AND-logic): $S = 3$: 4,193 county-years (CCD only; F-33 completeness below threshold); $S = 4$: 30,381 county-years (CCD + high F-33 completeness).

Sources: NCES CCD https://nces.ed.gov/ccd/files.asp; NCES F-33 https://nces.ed.gov/ccd/f33agency.asp. Scripts: `src/07_schools_nces_ccd.py`, `src/11_schools_f33.py`.

8.7. Local Finance

Primary sources: Census Annual Survey of State and Local Government Finances (ASGF), Census of Governments (CoG; census years 2012, 2017, 2022), and GSA Federal Audit Clearinghouse (FAC).

ASGF and CoG base structure. Both ASGF and CoG use the Census Government Master Address File (GoMAF) and collect individual government unit records that cross-reference to county FIPS, enabling county-level aggregation.

Census of Governments years ($S = 4$ for all counties): 2012, 2017, 2022. The Census of Governments is a complete enumeration of all government units, conducted every five years. CoG provides individual government unit records with full geocoding via the Government Master Address File, enabling sub-county geographic resolution (`spatial_grain` = 3). Combined with the Census Government timeseries API, $\leq$ 2-year freshness, and stable FIPS identifiers, CoG years satisfy all four $S = 4$ conditions for all counties.

ASGF survey years (2013–2016, 2018–2021): AND-logic applied. In ASGF survey years, individual government unit records are spatially aggregable only to the county level (`spatial_grain` = 2), so ASGF alone satisfies only the $S = 3$ condition.

The Federal Audit Clearinghouse (FAC) collects mandatory Single Audit submissions under OMB Uniform Guidance §200.501 from all entities expending $\geq$ \$750,000 in federal awards annually. Counties where at least one government entity files a Single Audit have an additional layer of mandatory, independently audited financial disclosure. The AND-logic constraint applies to ASGF years: counties *with* at least one FAC Single Audit filer retain $S = 4$; counties *without* a Single Audit filer are capped at $S = 3$.

FAC data are available via the GSA FAC API (https://www.fac.gov/api/) for fiscal years 2016 onward. ASGF years 2013–2015 pre-date the FAC API coverage window and score $S = 3$ for all counties. From 2016 onward, approximately 2,800–3,000 counties per year have at least one FAC filer.

Score distribution: $S = 3$: 10,727 county-years (ASGF years 2013–2015 for all 3,143 counties; ASGF years 2016, 2018–2021 for counties without FAC filer); $S = 4$: 23,847 county-years (all three CoG years; ASGF years with FAC filer from 2016 onward).

Sources: Census Government Finance https://www.census.gov/programs-surveys/gov-finances.html; Census of Governments https://www.census.gov/programs-surveys/cog.html; GSA Federal Audit Clearinghouse API https://www.fac.gov/api/. Scripts: `src/08_finance_census.py`, `src/11_finance_fac.py`.

8.8. Business and Labor

Primary sources: Census County Business Patterns (CBP) and BLS Quarterly Census of Employment and Wages (QCEW).

CBP base scoring and suppression proxy. CBP provides annual establishment counts, employment, and payroll by four-digit NAICS industry for all U.S. counties. The $S = 4$ criterion requires a suppression rate below 30% of four-digit NAICS cells, proxied by `ESTAB` $\geq$ 2,000. The 2019 threshold identifies **647 counties** ($\approx 20.6\%$) as low-suppression and upgrades them to $S = 4$ via the domain override in `11_score_domains.py`. This override is applied time-stably across all panel years.

QCEW uniformly scores $S = 3$ (`spatial_grain` $= 2$), so the CBP suppression override determines all $S = 4$ outcomes.

Score distribution: $S = 3$: 27,457 cells; $S = 4$: 7,117 cells.

Sources: Census CBP https://www.census.gov/programs-surveys/cbp.html; BLS QCEW https://www.bls.gov/cew/. Scripts: `src/09_business_cbp.py`, `src/10_business_qcew.py`.

9. Score Distributions and Cross-sectional Variation

Table 8 summarizes score distributions for each domain pooled across all 34,574 county×year observations ($D = 8$ core domains).

Active sources of cross-sectional variation. Several domains exhibit meaningful within-year cross-sectional variation. Table 9 reports the within-year coefficient of variation (σ/μ) for each domain, averaged only over *active years* — years in which at least some cross-county variation exists (i.e. $\sigma > 0$ and $\mu > 0$). Years in which all counties receive an identical score are excluded from the average.

1. **Health (5 active years):** In 2016–2020, 339 counties score $S = 4$ (500-Cities) while the remainder score $S = 3$, creating genuine cross-county spread (CV ≈ 0.083). Pre-2016 all counties score $S = 0$; from 2021 all score $S = 4$ — both phases contribute zero cross-sectional variation and are excluded from the active-year average.
2. **Environment:** Counties with EJScreen but no TRI or AQS monitor score $S = 3$ (≈ 300 counties/year); the remainder score $S = 4$ (from 2015); counties lacking all three score $S = 0$ (concentrated in 2012–2014).
3. **Housing:** Counties with a valid FHFA county HPI score $S = 4$ ($\approx 2{,}757$ per year); counties without HPI score $S = 3$ (≈ 387 per year).

4. **Transportation:** A graduated three-tier structure. Counties with NTD transit agencies score $S = 4$; counties with FARS crash records but no transit score $S = 3$; very remote counties with neither score $S = 0$.
5. **Schools:** Counties with CCD and high F-33 reporting completeness score $S = 4$ ($\approx$ 2,762/year); counties with CCD only score $S = 3$ ($\approx$ 381/year).
6. **Business/Labor:** 647 low-suppression counties score $S = 4$; the remaining 2,497 score $S = 3$ (time-stable).
7. **Local Finance (5 active years):** Cross-sectional variation exists only in ASGF years with FAC data (2016, 2018–2021). In those five years, counties with a FAC filer score $S = 4$ and counties without score $S = 3$. CoG years (2012, 2017, 2022) are uniform $S = 4$; ASGF years 2013–2015 are uniform $S = 3$ (pre-FAC API).
8. **Broadband:** Active-year variation is limited (4 active years, CV ≈ 0.008), concentrated in years where FCC+ACS qualifies most but not all counties for $S = 4$.

10. Variable Codebook

The final dataset contains 34,574 observations and 58 variables. Variables are organized in the order they appear in the dataset. Table 10 presents the complete codebook.

11. Summary Statistics

Table 11 presents summary statistics pooled across all 34,574 county×year observations.

Table 1: Source Registry Field Definitions (`config/sources_registry.csv`)

Field	Description
`source_id`	Unique snake_case identifier; used as the filename stem for evidence CSV outputs and referenced by `source_id` column in evidence files.
`project_role`	Classification: `core_domain` (receives a PDV score), `support` (feeds panel shell, controls, or outcomes but not scored).
`domain`	PDV domain label (health, environment, broadband, housing, transportation, schools, local_finance, business_labor) or foundation/controls/outcome for support sources.
`source_family`	Human-readable name of the data program.
`agency`	Responsible federal agency or organization.
`priority`	Implementation priority (1-high, 2-high, 3-medium, 4-hard).
`role_in_pdv`	One-sentence description of the source's role in the PDV scoring framework.
`official_homepage`	Canonical program landing page URL.
`api_or_download_url`	Primary bulk download or API endpoint.
`public_access`	Whether public access exists (yes/no/partial).
`account_or_key_required`	Whether an account or API key is needed.
`can_bypass_account`	Whether bulk downloads bypass account requirement.
`api_available`	Whether a documented REST or similar API exists.
`bulk_download_available`	Whether bulk file downloads at a stable URL are available.
`primary_geography`	The native geographic grain of the source (county, block group, facility, school, etc.).
`subcounty_geography`	Sub-county grain available (if any).
`years_needed`	Panel years for which data are needed.
`known_years_available`	Known data availability range.
`machine_formats`	Available machine-readable formats (CSV, JSON, shapefile, Excel, etc.).
`stable_identifiers`	Geographic or record identifiers present in the data (FIPS, GEOID, NCESSCH, etc.).
`download_strategy`	Brief operational note on how to collect the data (bulk download, API call, crosswalk, etc.).
`evidence_output`	Path to the evidence CSV this source produces (`data/interim/{source_id}_evidence.csv`).
`manual_audit_flag`	Whether the source requires manual audit or verification before use (yes/no).
`notes`	Free-text implementation notes and caveats.

Table 2: PDV Source Registry: Complete Catalog

Source ID	Domain	Agency	Primary grain	Access
Core domain sources (16 sources, 8 domains)				
`health_cdc_places`	Health	CDC	County / tract	Public; no key
`environment_ejscreen`	Environment	EPA	Block group	Public; API variable
`environment_tri`	Environment	EPA	Facility point	Public; no key
`environment_aqs`	Environment	EPA	Monitor point	Public; API key opt.
`broadband_fcc`	Broadband	FCC	Census block / address	Public; no key
`broadband_acs`	Broadband	Census Bureau	County	Public; key recommended
`housing_hud_chas`	Housing	HUD PD&R	County / tract	Public; API token for API
`housing_fhfa_hpi`	Housing	FHFA	County	Public; no key
`transportation_ntd`	Transportation	FTA	Transit agency	Public; no key
`transportation_nhtsa_fars`	Transportation	NHTSA	Crash point	Public; no key
`schools_nces_ccd`	Schools	NCES	School / LEA	Public; no key
`schools_nces_f33`	Schools	NCES	District (LEA)	Public; no key

Continued on next page

Table 2 continued

Source ID	Domain	Agency	Primary grain	Access
`finance_census_gov`	Local finance	Census Bureau	Government unit	Public; no key
`finance_fac`	Local finance	GSA / OMB	Government entity	Public; no key
`business_cbp`	Business/labor	Census Bureau	County	Public; key recommended
`business_qcew`	Business/labor	BLS	County	Public; no key
Support sources (non-scored; feed panel shell, controls, or outcomes)				
`support_county_geo`	Foundation	Census / MCDC / HUD	County / tract / ZCTA	Public; no key
`support_acs`	Controls	Census Bureau	County / tract / block group	Public; key recommended
`support_rucc`	Controls	USDA ERS	County	Public; no key
`support_usaspending`	Outcome	U.S. Treasury / OMB	Award / recipient	Public; no key

Table 3: PDV Scoring Rubric

Score	Label	Criterion
0	No data	No public record for county i in domain d exists in any repository in reference year t.
1	State-level only	Data exist but only as a state or higher-level aggregate; a county-specific estimate is unavailable.
2	County, stale or locked	County-level data exist but are updated less frequently than every 5 years, or are available only in formats requiring manual retrieval (locked PDF, request-only portals).
3	County, accessible	County-level data, updated within 5 years, available as bulk-downloadable machine-readable files (CSV, JSON, or shapefile) at a stable URL.
4	Sub-county, full access	Data at sub-county geographic grain (Census tract, block group, or finer); machine-readable; published via a documented REST API or FTP feed; updated within 2 years; records include stable FIPS or GEOID identifiers.

Table 4: PDV Composite Index and Sub-indices by Year (county-level means across 3,144 counties)

Year	PDV_raw	Std. dev.	Coverage	Resolution	Usability
2012	2.787	0.279	0.723	2.073	3.267
2013	2.661	0.282	0.722	1.947	3.264
2014	2.913	0.286	0.847	2.196	3.262
2015	3.010	0.191	0.866	2.415	3.340
2016	3.523	0.218	0.992	2.680	3.843
2017	3.658	0.206	0.991	2.803	3.964
2018	3.652	0.221	0.991	2.684	3.965
2019	3.650	0.217	0.993	2.686	3.970
2020	3.753	0.205	0.992	2.798	3.968
2021	3.756	0.199	0.992	2.798	3.969
2022	3.751	0.192	0.992	3.046	3.967
Pooled	3.374	0.474	0.918	2.557	3.707

Table 5: EPA AQS County Monitor Coverage by Year

Year	Counties with monitors	Share of 3,144
2012	1,094	34.8%
2013	1,084	34.5%
2014	1,078	34.3%
2015	1,079	34.3%
2016	1,069	34.0%
2017	1,074	34.2%
2018	1,072	34.1%
2019	1,066	33.9%
2020	1,051	33.4%
2021	1,046	33.3%
2022	1,044	33.2%

Table 6: HUD CHAS Vintage Mapping

Panel year	CHAS vintage	ACS end year
2012	2006–2010	2010
2013	2007–2011	2011
2014	2008–2012	2012
2015	2009–2013	2013
2016	2010–2014	2014
2017	2011–2015	2015
2018	2012–2016	2016
2019	2013–2017	2017
2020	2014–2018	2018
2021	2015–2019	2019
2022	2016–2020	2020

Table 7: NTD County Coverage by Year

Year	Counties with transit	Share of 3,144
2012	486	15.5%
2013	496	15.8%
2014	497	15.8%
2015	1,468	46.7%
2016	1,469	46.7%
2017	1,459	46.4%
2018	1,593	50.7%
2019	1,604	51.0%
2020	1,659	52.8%
2021	1,656	52.7%
2022	1,651	52.5%

Table 8: Domain Score Distributions ($N = 34{,}574$ county×year cells; $D = 8$ core domains)

Domain	$S = 0$	$S = 1$	$S = 2$	$S = 3$	$S = 4$	Mean	SD
Health	12,573	0	0	11,225	10,776	2.221	1.726
Environment	1,243	0	0	3,354	29,977	3.759	0.784
Broadband	0	6,287	0	9,443	18,844	3.181	1.113
Housing	0	0	0	4,244	30,330	3.877	0.328
Transportation	2,500	0	0	18,408	13,666	3.178	1.007
Schools	0	0	0	4,193	30,381	3.879	0.326
Local Finance	0	0	0	10,727	23,847	3.690	0.463
Business/Labor	0	0	0	27,457	7,117	3.206	0.404

Table 9: Within-year CV by domain (active years only; $D = 8$ core domains)

Domain	CV (σ/μ)	Active yrs / 11	Notes
Transportation	0.312	11/11	Three-tier $\{0, 3, 4\}$ structure every year
Environment	0.170	11/11	Monitor/TRI presence varies cross-sectionally
Business/Labor	0.126	11/11	CBP suppression threshold stable over time
Housing	0.085	11/11	FHFA HPI coverage gap in rural counties
Health	0.083	5/11	Only 2016–2020 active (partial PLACES rollout)
Schools	0.080	11/11	F-33 completeness gap in small districts
Local Finance	0.069	5/11	FAC filer vs. non-filer in ASGF years 2016, 2018–2021
Broadband	0.008	4/11	FCC+ACS $\rightarrow S = 4$; FCC alone $\rightarrow S = 3$; thin split

Table 10: Variable Codebook: All 58 Variables in `pdv_county_year.dta`

Variable	Type	Description
Identifiers (5 variables)		
`county_fips`	str	5-digit county FIPS code (zero-padded)
`state_fips`	str	2-digit state FIPS code (zero-padded)
`county_name`	str	County name (ACS 2020)
`year`	int	Calendar year (2012–2022)
`pop_2020`	int	County population, ACS 2020 5-year estimate
Composite index (3 variables)		
`PDV_raw`	float	Mean domain score: $(1/8)\sum_d S_{idt}$; range $[0, 4]$
`PDV_z`	float	Within-year z-score of `PDV_raw`
`PDV_pct`	float	Within-year percentile rank (0–100)
Sub-indices (9 variables)		
`PDV_coverage`	float	Share of 8 domains with $S_{idt} \geq 2$; range $[0, 1]$
`PDV_coverage_z`	float	Within-year z-score of `PDV_coverage`
`PDV_coverage_pct`	float	Within-year percentile rank of `PDV_coverage`
`PDV_resolution`	float	Mean R_{idt} across 8 domains; range $[0, 4]$
`PDV_resolution_z`	float	Within-year z-score of `PDV_resolution`
`PDV_resolution_pct`	float	Within-year percentile rank of `PDV_resolution`
`PDV_usability`	float	Mean U_{idt} across 8 domains; range $[0, 4]$
`PDV_usability_z`	float	Within-year z-score of `PDV_usability`
`PDV_usability_pct`	float	Within-year percentile rank of `PDV_usability`
Count (1 variable)		
`n_domains_scored`	int	Domains with $S_{idt} > 0$; range $[0, 8]$

Continued on next page

Table 10 continued

Variable	Type	Description
Domain scores (24 variables: base + _z + _pct for each of 8 domains)		
`score_health`	float	$\in \{0, 3, 4\}$; CDC PLACES
`score_health_z`	float	Within-year z-score of `score_health`
`score_health_pct`	float	Within-year percentile rank
`score_environment`	float	$\in \{0, 3, 4\}$; EJScreen + TRI + AQS (AND-logic)
`score_environment_z`	float	Within-year z-score
`score_environment_pct`	float	Within-year percentile rank
`score_broadband`	float	$\in \{1, 3, 4\}$; FCC + ACS B28002 (AND-logic)
`score_broadband_z`	float	Within-year z-score
`score_broadband_pct`	float	Within-year percentile rank
`score_housing`	float	$\in \{3, 4\}$; HUD CHAS + FHFA HPI (AND-logic)
`score_housing_z`	float	Within-year z-score
`score_housing_pct`	float	Within-year percentile rank
`score_transportation`	float	$\in \{0, 3, 4\}$; NHTSA FARS + FTA NTD
`score_transportation_z`	float	Within-year z-score
`score_transportation_pct`	float	Within-year percentile rank
`score_schools`	float	$\in \{3, 4\}$; NCES CCD + F-33 (AND-logic)
`score_schools_z`	float	Within-year z-score
`score_schools_pct`	float	Within-year percentile rank
`score_local_finance`	float	$\in \{3, 4\}$; Census ASGF/CoG + FAC (AND-logic)
`score_local_finance_z`	float	Within-year z-score
`score_local_finance_pct`	float	Within-year percentile rank
`score_business_labor`	float	$\in \{3, 4\}$; CBP + QCEW
`score_business_labor_z`	float	Within-year z-score
`score_business_labor_pct`	float	Within-year percentile rank
Resolution sub-scores R_{idt} (8 variables; spatial grain, 0–4)		

Continued on next page

Table 10 continued

Variable	Type	Description
`Ridt_health`	float	0 (2012–2015); 2–3 (2016+)
`Ridt_environment`	float	0 (2012–2014, no TRI); 3–4 (TRI/EJScreen/AQS)
`Ridt_broadband`	float	1 (2012–2013); 3 (2014–2021 FCC); 4 (2022 BDC)
`Ridt_housing`	float	3 all years (tract-level CHAS dominates)
`Ridt_transportation`	float	0 (no data); 2 (FARS only); 3 (NTD transit)
`Ridt_schools`	float	3 all years (school geocodes from CCD)
`Ridt_local_finance`	float	2 (ASGF years); 3 (CoG census years)
`Ridt_business_labor`	float	2 all years (county-level CBP/QCEW)
Usability sub-scores U_{idt} (8 variables; format access, 0–4)		
`Uidt_health`	float	0 (2012–2015); 3–4 (2016+)
`Uidt_environment`	float	0 (no data); 3–4 (TRI/EJScreen/AQS)
`Uidt_broadband`	float	1 (2012–2013); 3 (2014–2021); 4 (2022)
`Uidt_housing`	float	4 all years (CHAS API, bulk, 2-yr lag, FIPS)
`Uidt_transportation`	float	0 (no data); 3 (FARS or NTD bulk/FIPS)
`Uidt_schools`	float	4 all years (NCES API, bulk, 1-yr lag, FIPS)
`Uidt_local_finance`	float	3 (ASGF years); 4 (CoG: API, bulk, 2-yr lag, FIPS)
`Uidt_business_labor`	float	3 (high-suppression); 4 (low-suppression)

Table 11: Summary Statistics: PDV Panel, 2012–2022 ($N = 34{,}574$)

Variable	Mean	Std. dev.	Min	p25	Median	Max
Composite index and sub-indices						
PDV_raw	3.374	0.474	1.625	3.000	3.500	4.000
PDV_coverage	0.918	0.113	0.500	0.875	1.000	1.000
PDV_resolution	2.557	0.357	1.375	2.250	2.625	3.125
PDV_usability	3.707	0.364	2.375	3.375	4.000	4.000
Domain scores						
score_health	2.221	1.726	0	0.000	3.000	4
score_environment	3.759	0.784	0	4.000	4.000	4
score_broadband	3.181	1.113	1	3.000	4.000	4
score_housing	3.877	0.328	3	4.000	4.000	4
score_transportation	3.178	1.007	0	3.000	3.000	4
score_schools	3.879	0.326	3	4.000	4.000	4
score_local_finance	3.690	0.463	3	3.000	4.000	4
score_business_labor	3.206	0.404	3	3.000	3.000	4

Technical Appendix B:

Data Construction for the Competitive Federal Grant and Instrument Variables

Prepared for:

"The Legibility Premium: Public Data Visibility and the Allocation of Competitive Federal Grants"

May 2026

1 Overview

The paper's central claim is that a county's public data visibility shapes its capacity to capture federal grants *whose allocation depends on federal officers selecting among applicants.* Two empirical ingredients are required to test this claim: a measure of grant flow restricted to programs in which this selection mechanism operates, and plausibly exogenous shifters of public data visibility that can isolate the causal channel from confounding correlation.

The natural starting point for the outcome variable is the categorical classification field that the Federal Funding Accountability and Transparency Act (FFATA) schema applies to every federal financial assistance award. Awards classified as "Project Grants" or "Cooperative Agreements" under this schema are nominally the competitive categories. Two features of the underlying data make a direct equivalence between "classified as competitive" and "actually competitive" untenable. First, the agency classifying each award is the awarding agency itself, and agencies apply the FFATA categories inconsistently. The single largest program classified as a Project Grant in fiscal year 2020, accounting for more than one-fifth of dollar volume in the broad-classified bucket, is a formula-allocated transit grant. Other quasi-formula programs contribute substantial dollar volume to the same bucket. Second, the remaining truly competitive programs span agencies with very different selection mechanisms (peer review at NIH, panel review at NSF, discretionary selection at EDA), and treating them as homogeneous obscures the substantive mechanism the paper studies.

The competitive grants panel constructed here addresses both concerns by restricting the outcome variable to a hand-curated set of 34 federal grant programs in which (i) federal program officers, peer review panels, or interagency selection committees exercise substantial discretion in choosing among applicants; (ii) the program structure is unambiguously merit-based rather than formula-allocated; and (iii) the program has sustained dollar volume across the panel period.

The instrumental variables address a separate empirical concern. The within-county relationship between PDV and competitive grant capture might be confounded by reverse causality (federal grants generate reporting requirements that themselves produce publicly visible data) or by time-varying unobserved state capacity (state-level governance quality could drive both public data infrastructure and grant-writing competence). The IV strategy requires shifters of county-level PDV that are plausibly orthogonal to these confounding channels. State open-data laws and Chief

Data Officer positions provide one such source of variation: they shift state agency reporting practices in a way that propagates to county-level PDV but have no direct mandate to allocate federal grants. A Bartik shift-share predictor provides a second source of variation: it predicts county-level PDV changes from pre-period (2012) domain exposure interacted with subsequent national domain trends, with identification resting on the exogeneity of the pre-period exposure to the post-2012 federal data infrastructure expansions.

2 Defining the Program Universe

Three criteria, applied jointly, determine whether a Catalog of Federal Domestic Assistance (CFDA) program is included in the competitive grants panel.

Federal discretion in recipient selection. The first criterion is that award decisions are made by federal program officers, peer review panels, or interagency selection committees from a pool of applicants. Programs in which a statutory allocation formula mechanically determines which counties receive funds are excluded, regardless of how the awarding agency classifies them in the FFATA schema. The operational test is whether the program's CFDA listing at SAM.gov describes selection as "competitive," "project grant," or "cooperative agreement" and the underlying program documentation describes a peer-review or merit-review process.

Substantive competition. The second criterion is that the program receives more applications than it can fund and applicants must compete on the merit of proposals. Programs that effectively fund every eligible applicant — continuation awards without re-competition, formula-driven set-asides described as "competitive" by the awarding agency, and statutorily-mandated state-by-state allocations — are excluded. The operational test is whether the program's published award-rate documentation indicates that substantially more proposals are received than funded.

Sustained dollar volume. The third criterion is that the program disbursed at least $5 million nationally in a typical fiscal year between 2012 and 2022. This excludes pilot programs of short duration and one-off competitions whose inclusion would introduce panel composition shocks. The operational test is whether the program appears with non-trivial dollar volume in at least seven of

the eleven fiscal years in the panel.

The 34 programs admitted under these criteria span eleven federal agencies and represent three broad substantive areas. Research grants administered through peer review constitute the largest group, with NSF directorates and NIH institutes contributing twenty programs. Discretionary economic development and infrastructure grants from EDA, DOT, EPA, and FEMA contribute six programs. Health, education, and community programs from HRSA, SAMHSA, HUD, ED IES, USDA Rural Development, and DOJ contribute eight programs.

Table 1 presents the complete catalog.

The composition of the catalog reflects the substantive geography of discretionary federal grant-making. Research funding, dominated by NIH institutes and NSF directorates, accounts for the largest single block of programs because peer review is the canonical competitive selection mechanism in U.S. federal science policy. Economic development and infrastructure grants from EDA, DOT, EPA, and FEMA capture the major non-research discretionary streams aimed at place-based investment. The health, education, and community programs in the third group capture discretionary streams in which federal officers select among local applicants for substantive programmatic work.

3 Data Source and Aggregation

All award-level obligations are drawn from USAspending.gov, the public data portal mandated by the Federal Funding Accountability and Transparency Act. USAspending compiles transaction-level records of every federal financial assistance award; the records are published with substantial geographic and recipient detail and updated continuously. The panel uses the public REST API at `api.usaspending.gov`, specifically the geographic-aggregation endpoint that returns county-level totals filtered by program and fiscal year.

3.1 Geographic Attribution: Place of Performance

USAspending records two geographic attributions for each award. The *recipient location* field identifies the county in which the legal recipient entity is headquartered. The *primary place of performance* field identifies the county where the funded activity occurs. The two attributions diverge whenever the legal recipient and the locus of work are in different jurisdictions. Most consequentially, when a state-level entity (a state agency, a state university system, or a statewide nonprofit) receives a federal award on behalf of activities subsequently performed in specific counties, recipient location attributes the full obligation to the county in which the state recipient is headquartered — typically the state capital. Place of performance attributes the obligation to the county where the funded work actually occurs.

The panel uses place of performance throughout. The substantive question the paper studies is

where federal money is geographically directed for use, not where the legal grantee entity happens to be incorporated. Recipient-location attribution would conflate this question with the geographic clustering of state-government and large-nonprofit headquarters, biasing apparent grant capture toward state-capital counties and metropolitan centers in a way that has nothing to do with the substantive allocation of federal resources.

3.2 Query Structure

For each combination of program $p \in P$ (the 34 CFDA programs) and fiscal year $t \in \{2012, \ldots, 2022\}$, a single API request is issued. The request specifies place-of-performance geography at the county level, filters by the program's CFDA number, and restricts to obligations within the fiscal year defined as October of $t-1$ through September of t. The endpoint returns one record per county receiving any obligation under the specified program in the specified year. The complete pull comprises 374 API requests; all were executed in May 2026, and the cached JSON responses constitute the raw input to the panel construction.

3.3 County–Year Aggregation

For each county i and fiscal year t, the panel's level outcome is the sum of program-level obligations across the 34 programs:

$$G_{i,t} = \sum_{p \in P} g_{i,p,t}, \tag{1}$$

where $g_{i,p,t}$ is the place-of-performance obligation of program p to county i in fiscal year t. Counties with no obligation under any of the 34 programs in a given fiscal year are assigned $G_{i,t} = 0$. The zero designation is substantively important: a county that receives nothing from the competitive program universe is not equivalent to a county whose data are missing, and the analysis treats the absence of award as informative.

4 Variable Construction

The panel ultimately contains seven analytical variables per county– year cell, derived from the level obligation by a sequence of standardized transformations that align with the empirical speci-

fications.

4.1 Level and Real-Dollar Variables

The nominal-dollar obligation $G_{i,t}$ is the foundational level variable. Because the panel covers eleven fiscal years over which prices rose substantially, comparisons across years require deflation. The real-dollar series converts nominal dollars to constant 2022 dollars using the Bureau of Labor Statistics' Consumer Price Index for All Urban Consumers (CPI-U), annual averages:

$$G^{r}_{i,t} = G_{i,t} \cdot \frac{\mathrm{CPI}_{2022}}{\mathrm{CPI}_{t}}. \tag{2}$$

Table 1: The 34 Federal Grant Programs Comprising the Competitive Grants Panel

CFDA	Agency	Program
Research grants (peer-reviewed)		
47.041	NSF	Engineering
47.049	NSF	Mathematical and Physical Sciences
47.050	NSF	Geosciences
47.070	NSF	Computer and Information Science
47.074	NSF	Biological Sciences
47.075	NSF	Social, Behavioral and Economic Sciences
47.076	NSF	Education and Human Resources
47.078	NSF	Polar Programs
47.079	NSF	Office of International Science
47.083	NSF	Office of Integrative Activities
93.847	NIH/NIDDK	Diabetes, Digestive, and Kidney Diseases Research
93.853	NIH/NINDS	Neurosciences Research
93.855	NIH/NIAID	Allergy and Infectious Diseases Research
93.859	NIH/NIGMS	Biomedical Research and Research Training
93.866	NIH/NIA	Aging Research
93.273	NIH/NIAAA	Alcohol Research Programs
93.279	NIH/NIDA	Drug Abuse and Addiction Research
93.395	NIH/NCI	Cancer Treatment Research
93.396	NIH/NCI	Cancer Biology Research
93.398	NIH/NCI	Cancer Research Manpower
Economic development and infrastructure		
11.300	EDA	Investments for Public Works and Economic Development
11.307	EDA	Economic Adjustment Assistance
20.933	DOT	National Infrastructure Investments (BUILD / RAISE)
66.818	EPA	Brownfields Assessment and Cleanup Cooperative Agreements
66.469	EPA	Great Lakes Program
97.047	FEMA	Pre-Disaster Mitigation (BRIC)
Health, education, and community programs		
93.527	HRSA	New and Expanded Services under the Health Center Program
93.243	SAMHSA	Substance Abuse and Mental Health Services Projects of Regional and National Significance
10.351	USDA/RD	Rural Business Enterprise Grants
10.781	USDA/RD	Rural Cooperative Development Grants
14.273	HUD	Choice Neighborhoods Implementation Grants
84.305	ED/IES	Education Research, Development and Dissemination
84.215	ED	Promise Neighborhoods
16.812	DOJ	Second Chance Act Reentry Initiative

Table 2 reports the CPI-U values used.

Table 2: CPI-U Deflator (Annual Average, All Urban Consumers)

Fiscal Year	CPI-U	Fiscal Year	CPI-U
2012	229.594	2018	251.107
2013	232.957	2019	255.657
2014	236.736	2020	258.811
2015	237.017	2021	270.970
2016	240.007	2022	292.655
2017	245.120		

4.2 Per-Capita Transformations

The relevant economic outcome in cross-county comparison is grant capture per resident rather than absolute dollars, because the latter mechanically favors larger counties. Per-capita variables divide the level by the county's annual population estimate:

$$\widetilde{G}^{r}_{i,t} = G^{r}_{i,t} / N_{i,t}. \tag{3}$$

County population $N_{i,t}$ is the U.S. Census Bureau Population Estimates Program (PEP) annual estimate for July 1 of year t. For the small number of county–year cells with missing annual PEP estimates, the 2020 decennial Census count substitutes; the substitution affects fewer than 1% of cells.

4.3 Inverse Hyperbolic Sine Transformation

The headline regression outcome is the inverse hyperbolic sine of per-capita real-dollar obligations:

$$y_{i,t} = \operatorname{arcsinh}\big(\widetilde{G}^{r}_{i,t}\big) = \ln\Big(\widetilde{G}^{r}_{i,t} + \sqrt{{\widetilde{G}^{r}_{i,t}}^{2} + 1}\,\Big). \tag{4}$$

The transformation is selected because approximately 56% of county–year cells have $G_{i,t} = 0$. The logarithm would drop these observations entirely; $\log(1 + y)$ would handle them but with units-dependent distortion. The inverse hyperbolic sine is well-defined at zero, behaves like a logarithm for moderate-to-large values, and is invariant to the choice of monetary units. Coefficients in regressions where the IHS-transformed variable is the dependent variable are interpreted as

approximate proportional changes for the strictly positive portion of the distribution, with the approximation tightening as the value of the underlying variable increases. The relevant theoretical reference for this interpretation is Bellemare and Wichman (2020).

4.4 Extensive Margin and Program-Count Variables

Two auxiliary variables capture the extensive margin of grant receipt. A binary indicator equals one if the county received any competitive grant in the fiscal year:

$$\mathbf{1}\{G_{i,t} > 0\}. \tag{5}$$

A count variable records the number of distinct CFDA programs (out of 34) contributing to the county's total:

$$\sum_{p \in P} \mathbf{1}\{g_{i,p,t} > 0\}. \tag{6}$$

Both variables enter robustness regressions as alternative outcomes.

4.5 Share and Allocation-Ratio Variables

For framing grant allocation as a proportional question, two within-year share variables are computed. The first is the county's share of the national competitive-grant total in fiscal year t:

$$s_{i,t} = G_{i,t} / \sum_{j} G_{j,t}. \tag{7}$$

The second is the county's allocation ratio, defined as the grant share divided by the population share:

$$A_{i,t} = \frac{s_{i,t}}{N_{i,t} / \sum_{j} N_{j,t}}. \tag{8}$$

The allocation ratio centers on one for proportional allocation; values above one indicate that the county captures more grant dollars than its population share would imply, and values below one the reverse. The natural logarithm of the allocation ratio centers on zero and is the preferred form for share-based regressions.

5 Panel Structure

The constructed panel is a balanced 3,144 × 11 county–fiscal year panel that contains 34,574 observations before sample cleaning. After dropping county–year cells with missing identifiers or missing annual population (and a small number of state-aggregated rows that do not match the 5-digit FIPS structure), the analysis sample is 34,533 observations. County identifiers are 5-digit zero-padded FIPS codes; state identifiers are 2-digit FIPS. The panel covers all counties and county-equivalents in the 50 states and the District of Columbia for which a continuous PDV index is available over the period 2012–2022.

Fiscal year alignment. Federal fiscal years run from October of the prior calendar year through September. The year field in this panel is the fiscal year of obligation. The PDV index against which it is matched is computed at the calendar-year level. The merge between the two panels is contemporaneous in the baseline analysis. Because the timing of federal grant decisions typically precedes obligation by months to quarters, the analysis also reports specifications in which PDV is lagged by one year, aligning calendar year $t-1$ PDV with fiscal year t obligations.

Coverage. The panel contains observations for all 3,144 counties in all 11 fiscal years. Approximately 44% of county–year cells record positive competitive grant obligations; the remaining 56% record zero. The zero designation reflects substantive non-receipt rather than measurement absence: counties that did not win any of the 34 competitive programs in a given year truly received zero competitive grant dollars in that year. The empirical specification treats these zeros symmetrically with positive values through the inverse hyperbolic sine transformation.

6 Descriptive Statistics

This section reports descriptive moments of the competitive grants panel. The basic summary statistics in Table 3 establish the broad shape of the outcome variable; the analyses that follow document how the $281 billion in pooled obligations is distributed across agencies and programs, across geographic units, across recipient counties (concentration), and across time (persistence).

6.1 Aggregate Moments

Table 3: Summary Statistics for the Competitive Grants Panel ($N = 34{,}533$)

Variable	Mean	SD	Median	Max	Share > 0
Nominal-dollar obligation ($)	11.6M	102M	0	7.4B	0.44
Real-$2022 per-capita ($)	53.83	244	0	19,269	0.44
IHS of real-$2022 per-capita	1.44	2.41	0	10.55	—
Indicator: any competitive grant	0.44	0.50	0	1	—
Number of distinct programs (0–34)	2.11	4.73	0	30	—

The mean per-capita real-dollar obligation across all county–year cells is $54, but the distribution is heavily right-skewed: the median is zero (a majority of cells receive nothing in the narrow competitive universe), and the maximum is approximately $19,269 in a single county–year cell. The mean number of distinct programs contributing to each county–year cell is 2.1, with substantial right-skew up to a maximum of 30 of the 34 programs received by a single cell.

The annual time series, reported in Table 4, reveals substantial expansion in competitive-grant dollar volume over the panel period, with sharp increases beginning in 2014 and the largest single-year total recorded in 2020. The number of counties receiving any competitive grant rises monotonically from approximately 1,232 in 2012 to a peak of 1,517 in 2020 before declining to 1,454 in 2022. The drop in dollar volume between 2021 and 2022 reflects the phase-down of pandemic-era discretionary supplements rather than a structural contraction in the competitive program universe.

Table 4: Annual Competitive Grant Obligations, 2012–2022

Fiscal Year	Total Obligations ($B, nominal)	Counties with > 0
2012	18.3	1,232
2013	17.9	1,264
2014	22.7	1,287
2015	24.0	1,301
2016	25.2	1,340
2017	25.0	1,377
2018	27.0	1,403
2019	29.4	1,448
2020	34.4	1,517
2021	33.3	1,506
2022	24.2	1,454
Pooled total	**281.4**	—

6.2 Composition by Agency and Program

The 34 programs in the competitive grants panel are distributed unevenly in dollar terms across the eleven contributing federal agencies. Table 5 reports pooled obligations by agency over the full panel period. The National Institutes of Health alone contribute nearly half of all competitive grant dollars ($138.0 billion, 49.0% of the pooled total), reflecting the dominant role of biomedical research funding in the U.S. competitive grant landscape. The National Science Foundation contributes a further $75.8 billion (26.9%). The Health Resources and Services Administration's Health Center expansion program (CFDA 93.527), a single CFDA program, contributes $38.0 billion (13.5%). Together, these three federal funders account for nearly nine-tenths of all dollar volume in the competitive grants panel. The remaining eight agencies contribute the residual 10.4%.

Table 5: Pooled Obligations by Federal Agency, 2012–2022

Agency	Total Obligations ($B, nominal)	Share of Pooled Total (%)	Programs Contributing
NIH	138.0	49.0	10
NSF	75.8	26.9	10
HRSA	38.0	13.5	1
SAMHSA	11.0	3.9	1
EDA	7.3	2.6	2
DOT	4.7	1.7	1
ED (incl. IES)	3.3	1.2	2
EPA	1.4	0.5	2
FEMA	0.9	0.3	1
DOJ	0.6	0.2	1
USDA-RD	0.2	0.1	2
HUD	0.01	0.0	1
Total	**281.4**	**100.0**	**34**

Aggregated by substantive area, research grants account for 76.0% of the pooled total ($213.9 billion), health, education, and community programs for 18.9% ($53.2 billion), and economic development and infrastructure grants for 5.1% ($14.4 billion). The dominance of research funding within the curated competitive universe reflects the size of the NIH and NSF appropriations rather than a curation choice; the catalog admits a balanced set of agencies but cannot re-weight the underlying program-level appropriations.

The ten largest CFDA programs by pooled obligations are reported in Table 6. Eight of the ten are NIH or HRSA biomedical or health-services programs; the remaining two are NSF research

directorates (Mathematical and Physical Sciences, Education and Human Resources). The top three programs alone account for 34.1% of all dollar volume in the competitive grants panel.

Table 6: Top Ten Programs by Pooled Obligations, 2012–2022

CFDA	Agency	Program	$B	% of total
93.527	HRSA	New/Expanded Health Center Services	38.0	13.5
93.855	NIH/NIAID	Allergy and Infectious Diseases Research	33.4	11.9
93.859	NIH/NIGMS	Biomedical Research and Research Training	24.5	8.7
93.866	NIH/NIA	Aging Research	18.3	6.5
47.049	NSF	Mathematical and Physical Sciences	16.8	6.0
93.847	NIH/NIDDK	Diabetes, Digestive, and Kidney Diseases Research	16.5	5.9
93.853	NIH/NINDS	Neurosciences Research	15.9	5.6
47.076	NSF	Education and Human Resources	13.0	4.6
47.050	NSF	Geosciences	11.4	4.0
93.243	SAMHSA	Projects of Regional and National Significance	11.0	3.9
		Top 10 share of pooled total	—	**70.6**

6.3 Geographic Distribution

The geographic distribution of pooled obligations across U.S. states mirrors both the size and the research intensity of the underlying population. Table 7 reports the ten states with the largest pooled obligations over the panel period. California alone receives 13.7% of all dollar volume ($38.6 billion); together with New York (8.0%) and Massachusetts (7.4%), the three most research-intensive states capture 29.1% of pooled obligations. The top ten states capture 56.0%, and the bottom forty states and the District of Columbia divide the remaining 44.0%. The compositional geography is unsurprising: states with large NIH-funded medical research institutions and large NSF-funded research universities appear at the top of the distribution, while less research-intensive states with smaller populations appear at the bottom.

The geographic concentration at the state level does not, however, imply that competitive grants flow only to a narrow set of metropolitan recipients. The 2,530 counties that receive any competitive grant over the eleven-year panel are distributed across all fifty states and the District of Columbia. The state-level concentration documented above is driven primarily by the within-state concentration of grants in research-anchor counties (the counties containing major university and medical-school campuses) rather than by the geographic exclusion of entire states.

Table 7: Top Ten States by Pooled Obligations, 2012–2022

State	Total Obligations ($B)	Share of Pooled Total (%)
California	38.6	13.7
New York	22.4	8.0
Massachusetts	20.9	7.4
Texas	14.1	5.0
Pennsylvania	13.8	4.9
North Carolina	10.5	3.7
Illinois	10.4	3.7
Washington	9.7	3.5
Florida	8.8	3.1
Maryland	8.5	3.0
Top 10 share	—	**56.0**

6.4 Concentration and Inequality

The distribution of pooled obligations across counties is highly concentrated. Table 8 reports the share of the pooled total captured by the top n recipient counties for several values of n. The ten counties with the largest pooled obligations account for 27.2% of all dollar volume in the competitive grants panel; the top fifty counties account for 59.5%; and the top one hundred counties account for 75.1%. The implied Gini coefficient across all 3,144 counties (including the 614 counties with zero pooled obligations) is 0.928, indicating an extreme degree of geographic concentration. Even restricting attention to the 2,530 counties with positive pooled obligations, the Gini coefficient is 0.911.

Table 8: Concentration of Pooled Obligations across Counties, 2012–2022

Concentration measure	Share of pooled total (%)
Top 10 counties	27.2
Top 50 counties	59.5
Top 100 counties	75.1
Top 500 counties	94.0
Top 1,000 counties	98.5
Counties with >0 pooled obligations	2,530 of 3,144
Counties with zero pooled obligations	614 of 3,144
Gini coefficient (recipients only)	0.911
Gini coefficient (all 3,144 counties)	0.928

The concentration is substantively meaningful for the paper's empirical analysis. With three-quarters of pooled dollars flowing to the top one hundred counties, the cross-sectional variation in per-capita grant capture is dominated by the right tail of the distribution. The inverse hyper-

bolic sine transformation of the outcome variable, by compressing the right tail while preserving the substantive contrast between zero and positive cells, is well-suited to the empirical structure documented here.

6.5 Persistence in Grant Receipt

Whether a county receives competitive grants in one fiscal year is strongly predictive of whether it receives them in the next. Pooling across the ten consecutive-year transition pairs in the panel (2012–2013 through 2021–2022), the conditional probability that a county receives any competitive grant in year t given that it received one in year $t-1$ is 0.842. The conditional probability of receipt in year t given non-receipt in $t-1$ is 0.245. The implied stationary probability of receipt under the observed transition matrix is approximately 0.61, well above the within-year average of 0.44, reflecting the strong persistence in the receiving state.

The cumulative coverage across the eleven-year panel is a useful complement to the year-by-year statistics. Table 9 reports the distribution of counties by the number of fiscal years in which they received any competitive grant. Twenty-three percent of counties received a competitive grant in all eleven years of the panel; another sixteen percent received grants in nine or ten years. At the other extreme, only 1.7% of counties received nothing in any year of the panel — 54 counties total. The remaining 80% of counties fall in the middle: they received competitive grants in some years but not others, with the modal receiving-years count being one (12.6% of counties received in exactly one year).

Table 9: Distribution of Counties by Years of Positive Receipt, 2012–2022

Years with positive receipt	Counties	% of 3,144
Exactly 0	54	1.7
Exactly 1	396	12.6
Exactly 2–3	484	15.4
Exactly 4–6	423	13.5
Exactly 7–8	267	8.5
Exactly 9–10	945	30.1
Exactly 11	575	18.3
At least one year	2,476	98.3
All eleven years	575	18.3

The dual pattern — strong year-to-year persistence among recipients combined with the exis-

tence of intermittent recipients and a small but non-zero share of always-non-recipient counties — is consistent with the paper's interpretation of a visibility-based screening process. Recipients accumulate documentary capacity through the act of receiving and reporting on awards, raising their visibility and hence their probability of receiving the next round of competition. Non-recipients, lacking the documentary infrastructure that participation generates, face a structurally lower probability of crossing the visibility threshold in subsequent years. The persistence patterns documented here are themselves a downstream consequence of the mechanism the paper studies.

7 Construction of the Instrumental Variables

The empirical analysis employs three instrumental variables to address the endogeneity of public data visibility in the within-county panel specification. The state open-data law indicator and the state Chief Data Officer indicator together constitute the "state-level policy" instrument set used in the IV columns of the main paper. The Bartik shift-share predictor constitutes the "shift-share" instrument used as an alternative identification strategy. This section documents the coding of each variable and discusses the identifying assumptions.

7.1 Why Instrumental Variables Are Needed

The two-way fixed effects regression in the main paper estimates the within-county relationship between PDV percentile rank and competitive grant capture, with county fixed effects absorbing all time-invariant unobservables and state-by-year fixed effects absorbing state-level shocks. Two endogeneity concerns remain.

The first is reverse causality. Federal grants come with reporting requirements: agencies that disburse funds also require recipients to report on use of funds, which generates publicly visible data documenting the recipient county's conditions and outcomes. A county that wins a federal grant in year $t-1$ may therefore appear more visible in year t in part *because of* the grant rather than in advance of it. The within-county TWFE coefficient is biased by this channel; the direction of the bias depends on the strength of the feedback relative to the direct effect of visibility on grant selection.

The second is time-varying unobserved state capacity. State-level governance quality may si-

multaneously drive state agency data-publication practices (raising county PDV) and the technical quality of grant applications submitted by counties in the state (raising grant capture). Both effects could be unobserved in the panel and would generate a positive within-county correlation between PDV and grants that does not reflect a causal effect.

The instrumental variables strategies isolate variation in PDV that is plausibly independent of these confounding channels.

7.2 State Open-Data Law Indicator

The first instrument, $Z^{\text{ODL}}_{s(i),t}$, is a binary indicator equal to one if county i's state has adopted a state-level open-data law or policy by year t. State open-data laws mandate that state agencies publish administrative data in machine-readable form, often with specifications regarding format, frequency, and accessibility. These mandates apply to state agency reporting and propagate to county-level PDV by making state administrative records relevant to county-level outcomes publicly visible.

Coding rule. A state is coded as having an open-data law in effect in year t if either of the following is true as of the start of fiscal year t: (i) the state has enacted a statute mandating that designated state agencies publish administrative data in machine-readable, publicly accessible form; or (ii) the state's governor has issued an executive order to the same effect. The variable takes value 1 from the year of adoption forward and 0 in earlier years. States that have not adopted such a policy by 2022 take value 0 throughout the panel.

Sources. Adoption dates are coded from a combination of three sources. The primary source is the National Conference of State Legislatures' state-by-state tracker of open-data legislation, which catalogs state statutes by year. The second is a hand-compiled catalog of state executive orders relevant to open data, drawn from individual state government websites and the National Association of State Chief Information Officers' state policy archive. The third is the Public Health Law Center's catalog of state open-data initiatives, used to cross-validate dates and identify policies not captured by the first two sources. Where multiple instruments establish open-data obligations, the earliest binding instrument is used as the adoption date.

Coverage. By 2022, twenty-one states have an open-data law or executive order in effect under this coding. The earliest adopters are New York (2013, executive order; 2017, statute), California (2014, executive order; 2016, statute), and Illinois (2014, executive order). The distribution of adoption dates is right-skewed within the panel: the modal adoption year is 2017–2019. States that have not adopted by 2022 include most of the South and parts of the Mountain West.

Identifying assumption. The exclusion restriction is that state open-data laws affect competitive grant capture only through their effect on county-level PDV, conditional on county fixed effects and year (or state-by-year) fixed effects. This is defensible on substantive grounds because state open-data laws apply to state agencies rather than to federal grant allocation, and they do not contain provisions that directly favor counties in the adopting state for federal grant capture. The first-stage relationship between the instrument and PDV is reported in the main paper: the coefficient in the single-instrument first stage is 2.0028 with a Kleibergen-Paap weak-IV F-statistic of 22.37, substantially above conventional weak-IV thresholds.

7.3 State Chief Data Officer Indicator

The second instrument, $Z^{\text{CDO}}_{s(i),t}$, is a binary indicator equal to one if county i's state has created a Chief Data Officer position by year t. State Chief Data Officers are executive-branch positions responsible for state government data infrastructure: coordinating data publication across agencies, developing standards for machine-readable reporting, and overseeing public access to state administrative records. The creation of a CDO position shifts state agency data practices in a manner analogous to but distinct from open-data law adoption.

Coding rule. A state is coded as having a CDO in effect in year t if a formal state-level Chief Data Officer position has been created by either statute, executive order, or formal departmental appointment as of the start of fiscal year t. The variable takes value 1 from the year of creation forward and 0 in earlier years.

Source. The primary source is the Beeck Center for Social Impact and Innovation's State CDO Tracker maintained at Georgetown University, which catalogs state CDO positions by year of establishment and provides documentation on the formal mechanism of creation. The tracker is

cross-validated against state government press releases and the National Association of State Chief Information Officers' records.

Coverage. By 2022, twenty-eight states have created a Chief Data Officer position under this coding. The earliest adopters are Colorado (2014), New York (2016), and Indiana (2016). The instrument is distinct from but correlated with the open-data law indicator: states that adopt one policy are more likely to adopt the other, but the timing and substantive content of the two policies differ substantially. The first-stage relationship between the CDO indicator and PDV is reported in the main paper: the coefficient in the two-instrument first stage is 1.5199, with the joint first-stage F-statistic equal to 15.13.

Identifying assumption. The exclusion restriction is parallel to that for the open-data law instrument. CDO positions shift state agency data practices but have no direct role in federal grant allocation. The two instruments together provide overidentification: tests of overidentifying restrictions can be conducted using the two-instrument specification.

7.4 Bartik Shift-Share Predictor

The third instrument, $Z^{\mathrm{B}}_{i,t}$, is a continuous Bartik shift-share predictor constructed from county-level baseline (2012) domain exposure interacted with national domain trends through year t. The instrument exploits the fact that counties with greater pre-period exposure to PDV domains that subsequently grew nationally experience larger predicted PDV increases, with the cross-sectional exposure pattern fixed in 2012 (before the federal data infrastructure expansions of the panel period) and the national trends taken as the "shift" component.

Construction. For each county i in 2012, the baseline share of its composite PDV attributable to domain d is computed as

$$\omega_{i,d,2012} = \frac{S_{i,d,2012}}{\sum_{d'=1}^{8} S_{i,d',2012}}, \tag{9}$$

where $S_{i,d,2012}$ is the raw 0–4 score on the PDV rubric in domain d for county i in 2012. The denominator normalizes the shares so that $\sum_d \omega_{i,d,2012} = 1$ for every county. For each domain d

and year t, the national mean of the domain score is computed as

$$\overline{S}_{d,t} = \frac{1}{N} \sum_{j=1}^{N} S_{j,d,t}, \tag{10}$$

where $N = 3{,}144$ is the number of counties in the panel. The shift component for domain d between 2012 and year t is the national mean difference $\overline{S}_{d,t} - \overline{S}_{d,2012}$. The Bartik predictor is the inner product of baseline shares and national shifts:

$$Z_{i,t}^{\mathrm{B}} = \sum_{d=1}^{8} \omega_{i,d,2012} \cdot \left(\overline{S}_{d,t} - \overline{S}_{d,2012} \right). \tag{11}$$

Identifying assumption. Identification follows the framework of Goldsmith-Pinkham et al. (2020): the exclusion restriction reduces to a requirement that the baseline 2012 domain shares $\omega_{i,d,2012}$ are exogenous with respect to the post-2012 federal data infrastructure expansions that drive the national trend component $\overline{S}_{d,t} - \overline{S}_{d,2012}$. The 2012 cross-sectional pattern of domain visibility predates the major federal data infrastructure expansions of the panel period: the launch of EJScreen (2015), the nationwide expansion of the CDC PLACES program (2020), the FCC Broadband Data Collection rebuild (2022), and others. The pre-period exposure pattern is therefore plausibly orthogonal to the within-panel evolution of national domain trends.

First-stage diagnostics. The first-stage regression of PDV percentile rank on the Bartik predictor yields a coefficient of 38.4272 with a Kleibergen-Paap weak-IV F-statistic of 54.26. The Bartik predictor is a substantially stronger instrument than the state-level policy indicators on conventional weak-IV criteria, and the second-stage coefficient of 0.0775 from the Bartik IV specification is reported in the main paper.

7.5 Coordination Across Instruments

The three instruments exploit different sources of variation in county PDV: the open-data law indicator captures shifts in state agency reporting practices induced by statutory mandate; the CDO indicator captures shifts induced by executive-branch coordination; and the Bartik predictor captures the interaction between fixed county-level exposure and time-varying national trends.

The independence of these sources allows the main paper to report each instrument separately as a cross-validating check on the IV identification strategy. The coefficient on PDV is positive and statistically significant across all three IV specifications reported in the main paper, and the magnitudes are broadly comparable to within an order of magnitude, suggesting that the underlying causal relationship is robust to the choice of identification strategy.

8 Limitations

Several limitations of the constructed objects are worth disclosing.

Program-list curation. The 34-program list is curated by the author. The selection criteria of Section 2 are explicit, the program list is documented, and the resulting catalog of programs spans the major agencies and substantive areas of competitive federal grant-making. Nevertheless, reasonable researchers may differ on individual program inclusion. The main paper reports leave-one-out sensitivity that drops each program in turn and re-estimates the headline regression; results are not driven by any single program.

Place-of-performance resolution. Place-of-performance attribution, while substantively preferable for the paper's research question, is not always county-resolved in USAspending. For approximately 3% of dollar volume in the panel, the place-of-performance field is recorded only at the state level and does not contribute to the county-level totals. This affects the level of obligations recorded for some counties but does not introduce systematic bias in the relationship between PDV and grant capture as long as the unresolved geography does not correlate with PDV.

Fiscal-year alignment. Federal fiscal years run from October of the prior calendar year through September. The PDV index is calendar-year. The contemporaneous merge introduces a partial timing overlap; the lagged specifications in the main paper address this.

Open-data law coding. The state-level open-data law instrument is hand-coded from public sources. Reasonable researchers may differ on whether particular state executive orders, departmental policies, or statutes qualify as "open-data law adoption." The coding rule documented

above prioritizes statutes and gubernatorial executive orders with explicit machine-readable publication mandates, but borderline cases (informal policies, departmental open-data programs without formal mandate) are omitted. Robustness to alternative coding decisions can be tested by varying the inclusion rules.

Bartik baseline shares. The Bartik predictor uses 2012 as the baseline year. Alternative baseline years (2010, 2008, or an average over a pre-period window) would generate slightly different predictors. The 2012 baseline is selected because it is the first year of the panel and immediately precedes the major federal data infrastructure expansions of the post-2012 period; this is the cleanest pre-period for the identification argument. Sensitivity to alternative baseline years could be reported in a robustness appendix.

9 Reproducibility

The complete pull and construction pipeline is included in the replication package. The Python program that issues the 374 API requests for the competitive grants panel is idempotent: re-running it skips cached program–year responses. The list of 34 programs and the metadata documenting agency, substantive area, and selection criteria is stored separately as a machine-readable specification file. The Stata construction script takes the cached API responses, joins them to the county roster for population and CPI deflator merge, applies the per-capita and inverse hyperbolic sine transformations documented above, and writes the final panel. The hand-coded state open-data law and Chief Data Officer indicators are stored as a state-year panel that merges 1:1 with the main analysis file. The Bartik predictor is constructed entirely from the PDV index and requires no external data inputs beyond the panel itself; the construction is implemented in Stata as part of the main analysis file. The complete pipeline runs in approximately fifteen minutes on a standard laptop with API access.

References

Bellemare, M.F. and C.J. Wichman (2020). "Elasticities and the Inverse Hyperbolic Sine Transformation." *Oxford Bulletin of Economics and Statistics*, 82(1): 50–61.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020). "Bartik Instruments: What, When, Why, and How." *American Economic Review*, 110(8): 2586–2624.