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 **(author?)** (1).

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:

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

Identifying assumption. Identification follows the framework of **(author?)** (2): 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

[1] 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.

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