Federal government wage indexes

(Official version at https://www.bls.gov/opub/mlr)

Travis Cyronek and Ted To

Office of Compensation and Working Conditions
US Bureau of Labor Statistics

April 2023

For nearly 50 years, the Employment Cost Index (ECI) has been providing the public with estimates of the change in employer labor costs. We explore the practicality of constructing federal wage indexes, in the spirit of the ECI, using Office of Personnel Management (OPM) salary data. To accomplish this task, we aggregate OPM records into occupation and industry groups. Although these salary data have a crosswalk for mapping OPM occupation codes into the Standard Occupational Classification system, no corresponding crosswalk exists for industries. A key hurdle, therefore, involves creating a crosswalk that assigns industry codes to OPM establishments. We create this crosswalk by developing an algorithm that uses Quarterly Census of Employment and Wages data and machine-learning tools to match agencies with a unique industry. With this agency-North American Industry Classification System crosswalk, we calculate annual Laspeyres, Paasche, and Fisher wage indexes for several aggregations. The resulting wage inflation rates are plausible and do not deviate substantially from the corresponding private industry and state and local wage inflation rates.

The Employer Cost Index (ECI) of the National Compensation Survey (NCS) has provided the public with estimates of changes in labor costs since December 1975. At the ECI launch, only private industry estimates were published; however, in June 1981, ECI expanded to include state and local

government workers. The federal government, despite being the largest U.S. employer with over 3 million employees (see table 1), is presently out of scope for NCS data products. This article explores, as a proof of concept, the practicality of constructing federal wage indexes using Office of Personnel Management (OPM) salary data. Since this analysis is purely exploratory, we do not attempt to fully replicate ECI methodology, but instead use it as a guide.

To construct federal wage indexes, we must overcome one major hurdle: records from the OPM data must be categorized into industry (see appendix table A-1) and occupation groups (see appendix table A-2) that are consistent with NCS aggregations used for the ECI.[1] The latter is straightforward because the U.S. Bureau of Labor Statistics (BLS) uses a crosswalk classification system to map OPM occupations into the Standard Occupational Classification (SOC) system. The former, in contrast, is more difficult because the OPM data do not contain industry codes. To address this problem, we use the department and agency information in the OPM data and machine-learning tools to match OPM and Quarterly Census of Employment and Wages (QCEW) establishments.[2] An algorithm is developed to select a unique North American Industry Classification System (NAICS) code for each agency observed in the OPM data. This final mapping yields a desired agency-to-NAICS crosswalk that we use to calculate Laspeyres, Paasche, and Fisher wage indexes for a variety of aggregations.[3]

Wage index number formulas

We have many index number formulas to choose from, including the commonly used Laspeyres and Paasche indexes and the less commonly used Dutot or Jevons indexes.[4] For exploratory purposes and brevity, we focus on the Laspeyres, Paasche, and Fisher indexes.

Given wages and employment for periods 0 (base period) and 1 (comparison period), the Laspeyres and Paasche wage index number formulas use a fixed "basket" of jobs (employment) to compute the ratio of total wage costs for period 1 to total wage costs for period 0. The Laspeyres index uses the fixed basket to be period-0 employment, whereas the Paasche index uses the

Table 1: Summary Statistics

	2020Q2		2021Q2		2022Q2	
	N	**%**	**N**	**%**	**N**	**%**
PSOC						
Management, business, and financial occupations	1,588,381	0.50	1,608,050	0.49	1,604,617	0.50
Professional and related occupations	924,123	0.29	949,506	0.29	940,476	0.29
Office and administrative support occupations	301,965	0.09	304,487	0.09	293,667	0.09
Service occupations	251,655	0.08	255,296	0.08	244,730	0.08
Transportation and material moving occupations	75,143	0.02	74,427	0.02	71,750	0.02
Construction, extraction, farming, fishing and ...	17,075	0.01	16,680	0.01	16,408	0.01
Installation, maintenance and repair occupations	18,298	0.01	18,862	0.01	18,380	0.01
Production occupations	12,549	0.00	12,496	0.00	12,115	0.00
Sales and related occupations	10,908	0.00	10,279	0.00	9,330	0.00
PNAICS						
Public administration	3,005,275	0.94	3,052,558	0.94	3,005,394	0.94
Rest of Services	80,381	0.03	82,526	0.03	83,502	0.03
Hospitals	39,908	0.01	40,945	0.01	49,224	0.02
Wholesale and Retail Trade	37,199	0.01	36,203	0.01	35,752	0.01
Goods Producing	1,1076	0.00	11,176	0.00	11,205	0.00
Elementary and secondary schools	9,012	0.00	9,176	0.00	9,299	0.00
Transportation and warehousing	7,123	0.00	7,117	0.00	7,345	0.00
Rest of Health Services	5,975	0.00	6,174	0.00	6,126	0.00
Colleges, universities, and professional schools	3,273	0.00	3,330	0.00	2,774	0 00
Nursing and residential care facilities	875	0.00	878	0.00	852	0.00
Full/Part-time						
full-time	3,097,080	0.97	3,147,790	0.97	3,115,654	0.97
part-time	103,017	0.03	102,293	0.03	95,819	0.03
Total	3,200,097	1.00	3,250,083	1.00	3,211,473	1.00

Source: Author's calculations using data from the Office of Personnel Management.

fixed basket to be period-1 employment. These formulas are given by

$$I_L = \sum_{i=1}^{n} s_i^0 \frac{w_i^1}{w_i^0} \tag{1}$$

and

$$I_P = \left(\sum_{i=1}^{n} s_i^1 \left(\frac{w_i^1}{w_i^0} \right)^{-1} \right)^{-1} \tag{2}$$

where I_L and I_P are the Laspeyres and Paasche indexes, w_i^t is hourly wage, s_i^t is the expenditure share, and i is job $1, 2, \ldots, n$. The expenditure share is given by

$$s_i^t = \frac{w_i^t e_i^t}{\sum_{j=1}^{n} w_j^t e_j^t}, \tag{3}$$

where e_i^t is employment, i and j are jobs $1, 2, \ldots, n$, and t is period $0, 1$.[5] In theory, employers can be expected to substitute away from more expensive workers. Since the Laspeyres index uses a period-0 fixed-employment basket, the Laspeyres index theoretically overstates wage inflation. Conversely, since the Paasche index uses a period-1 fixed-employment basket, the Paasche index theoretically understates wage inflation.

The Fisher wage index is given by the geometric mean of the Laspeyres and Paasche indexes as

$$I_F = \sqrt{I_L I_P}. \tag{4}$$

Along with the Törnqvist index, the Fisher index is considered to be "superlative," with a base and comparison period treated symmetrically to better capture labor substitution effects.[6]

Data

BLS has four quarters of OPM data: first quarter of 2019 and second quarter of 2020, 2021, and 2022. For this analysis, we omit the data from the first quarter of 2019 for two reasons. First, 2019 (first quarter) to 2020 (second quarter) straddled the start of the COVID-19 pandemic, which saw large and uncharacteristic changes in the labor market. Second, 2019 (first quarter) to 2020 (second quarter) was a five-quarter period that included two federal

salary increases. The data cover workers employed at the end of each quarter. Note that the data are reported to OPM by human resource offices across the federal government and may be subject to some error. If the federal workforce were incorporated into the ECI, data would need to be collected quarterly from OPM.

OPM data include individual federal employees, annual salary, OPM occupation, full-time or part-time status,[7] grade, agency, city, and state. BLS's OPM data include workers on military bases (which we exclude) but not postal service employees.[8] These data do not include any benefit-cost data (e.g., health insurance, retirement, nonproduction bonuses). All salaries are given as annual full-time salaries, so hourly wages are computed by dividing salary by 2,087.[9] Missing from OPM data are industry data (NAICS codes), so we use QCEW data and some machine-learning tools to construct an agency-to-NAICS concordance.

Also missing from the OPM data are establishment identifiers. So, we identify them by what we observe: agency, city, and state data, which can be used as imperfect proxies for an establishment. When an agency has just a single establishment within a city, city and state work as a perfect proxy. But if an agency has multiple establishments within a city, city and state are imperfect because multiple establishments are identified as a single establishment.

With an agency-to-NAICS crosswalk and a method for identifying establishments, we then map SOC and NAICS codes into occupation and industry groups (sometimes referred to as pseudo-SOC [PSOC] and pseudo-NAICS [PNAICS]). (See appendix tables A-1 and A-2.) Mean wages and total employment are computed for each basic ECI cell (a grouping by PSOC, PNAICS, and job) or subcell (a grouping by PSOC, PNAICS, subcell category, and job). Summary statistics, including employment counts and percentages of total employment from the OPM, are presented in table 1.

Since this analysis is purely exploratory, we do not attempt to reproduce the method for computing the ECI but instead use its basic conceptual framework for computing wage cost indexes for common index number formulas.[10] For the ECI, the unit of observation is a quote (such as an establishment, occupation, work status, or grade). These quotes are aggregated into cells consisting of an ownership sector, industry group (PNAICS), and occupation

group (PSOC). Cells can be further divided into subcells that may include full- or part-time status, region, division, union status, and so forth.

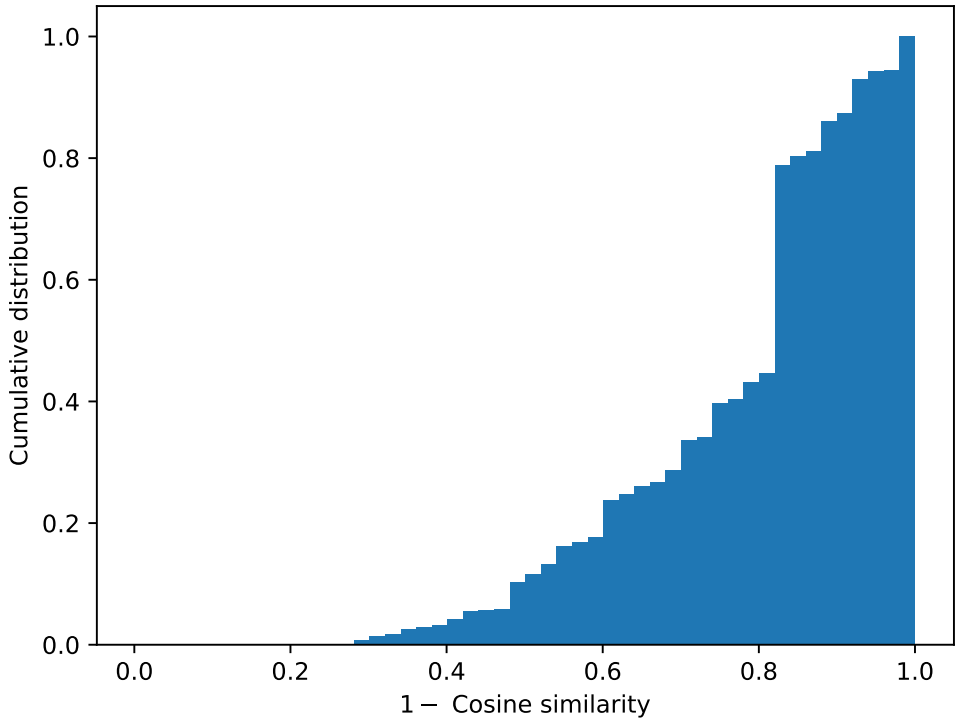
NAICS codes

Missing from the OPM data are NAICS codes. We construct an agency-to-NAICS crosswalk using QCEW-reported NAICS codes for federal government establishments. The OPM data have standardized, descriptive text for each department and agency. In the QCEW, the department, agency, and NAICS codes are reported individually by each establishment. These reports are subject to variations in establishment practice and can include spelling errors and varying abbreviations. For these reasons, matching the OPM establishments with QCEW establishments is not straightforward.

To construct an agency-to-NAICS crosswalk, we begin by aggregating individual employee data in the OPM data to agency by location. We then match each OPM agency and location with each QCEW establishment by year or quarter, state, and county. For each of these matches, cosine similarities are then calculated for term frequency–inverse document frequency (or TF–IDF) vectorized department descriptions and agency descriptions. This approach essentially amounts to the construction of a cardinal measure of similarity between two vectors. A number of options exist for constructing these vectors for a given match's descriptions. We have explored bag-of-words unigrams (an unordered list of the individual words from the descriptions) and character n-grams (a contiguous sequence of n characters from a piece of text). We ultimately chose character n-grams because they account for the issue of spelling errors or variations. A key problem with selecting a vectorization strategy is the lack of an objective standard. That is, in the absence of an objective standard, any choice between vectorization strategies possesses some level of arbitrariness.

For a given vectorizer, we use the mean of the cosine similarities for department and agency, weighting by QCEW-reported mean employment and upweighting and downweighting by the relative deviation between employee counts in the OPM establishment-level data and QCEW-reported mean employment. We assume here that larger establishments are more reliable but may also be "punished" for large differences in the reporting of a variable that should be similar. The QCEW department or agency with the best weighted cosine similarity is chosen as the match.

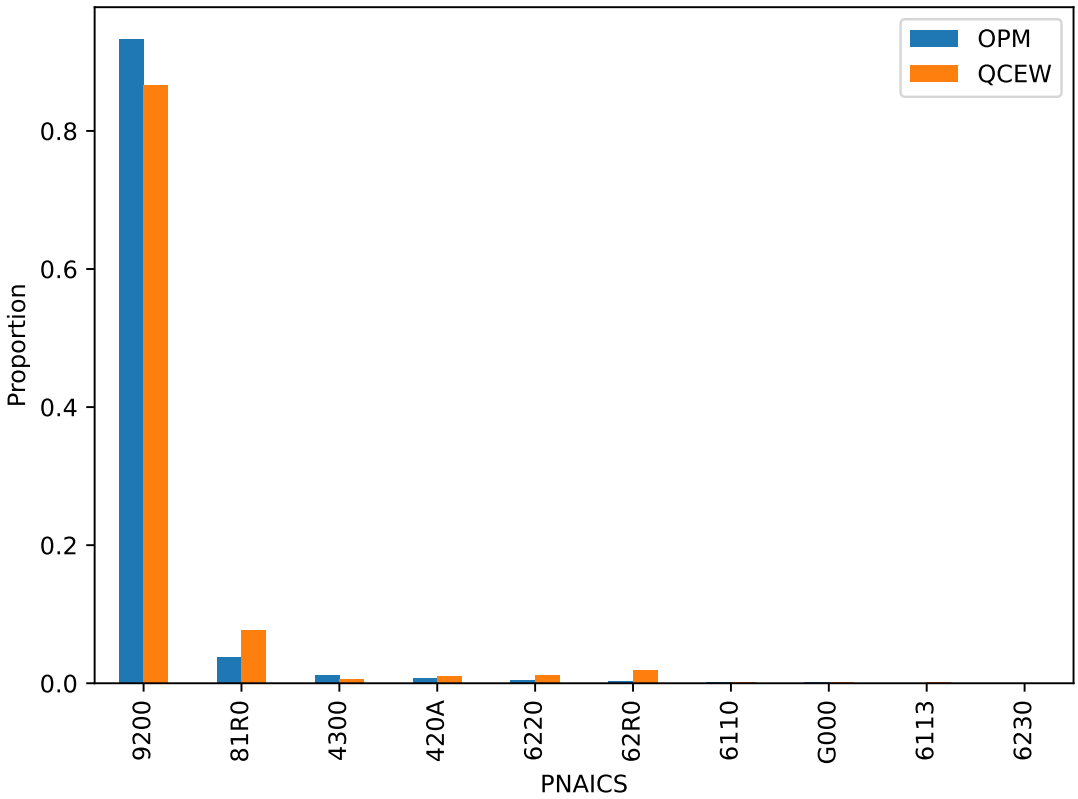
Chart 1: Distribution of cosine similarity scores for selected Quarterly Census of Employment and Wages matches, weighted by agency size



Source: Author's calculations.

Finally, since each department or agency should uniquely match a NAICS code, we compare the weighted cosine similarity among all establishments for a department or agency and select the NAICS code for the establishment with the best matching weighted cosine similarity. As constructed, the crosswalk is not without flaws, with a mean agency-size weighted score of 0.76 (standard deviation 0.161) and ranging from nearly the worst (0.002) to the best (1.000). The cumulative distribution of cosine similarity scores, weighted by agency size (see chart 1), shows that the bulk of matches are fairly reliable (> 0.8), with very few that are clearly unreliable (< 0.4). Moreover, the federal government distribution of PNAICS in the OPM dataset roughly matches that for the QCEW data (see chart 2).

Chart 2: Comparison of the distribution of PNAICS codes with OPM and QCEW data, second quarter of 2020



Note: OPM = Office of Personnel Management, PNAICS = pseudo-North American Industry Classification System, and QCEW = Quarterly Census of Employment and Wages.
Source: Author's calculations.

For computing exploratory wage indexes, this imperfect crosswalk is sufficient. But to publish indexes using OPM data will require dedicated analyst labor to create a more accurate crosswalk.

Wage index calculations

To compute wage indexes, we first partition the OPM microdata into establishments (department, agency, and city and state) and jobs (occupation, full- or part-time status, and grade).[11] Next, we compute average hourly rates and number of employees for each job within an establishment. The establishment-job data are then matched between the second quarter of 2020 and the second quarter of 2021 and between the second quarter of 2021 and the second quarter of 2022. The resulting matched data are partitioned by cell (PNAICS and PSOC) and period. We then calculate weighted average wages and total employment. Finally, we aggregate these data into wage indexes with the use of the Laspeyres, Paasche, and Fisher formulas. To compute subcell wage indexes, we partition the matched establishment-job data by subcell (PNAICS, PSOC, subcell category). Then, we calculate weighted average wages and total employment and aggregate them into subcell wage indexes. Note that for the published ECI, the base period is fixed and all comparisons are relative to the current base quarter (currently the fourth quarter of 2005). In contrast, for each matched pair of OPM datasets (e.g., the first quarter of 2020 to the second quarter of 2021), the base period is the earlier time (e.g., the first quarter of 2020) so that the time series of indexes for each cell and subcell is what is termed "chained."

Laspeyres, Paasche, and Fisher wage index calculations are shown in tables 2 through 6 for the basic cell aggregation and for a variety of subcell aggregations. We find that our computed rates of inflation are reasonable. Note that the calculations of the Laspeyres, Paasche, and Fisher wage indexes are quite close and, in some instances, equal up to the fourth decimal. This result is similar to other research results.[12] This present research also showed that the expected pattern in which the Laspeyres index exceeds the Paasche index is frequently reversed.[13] Finally, a comparison of the federal Laspeyres index with the official ECI is given in table 7. Perhaps unsurprisingly, the exploratory federal ECI is more closely aligned with the state and local ECI.

Table 2: Wage index calculations of basic cell, 2020 second quarter to 2022 second quarter

Period	**Laspeyres**	**Paasche**	**Fisher**
2020 Q2 to 2021 Q2	1.0131	1.0131	1.0131
2021 Q2 to 2022 Q2	1.0342	1.0341	1.0341

Note: Q2 = second quarter. Wage index data are aggregated into basic cells consisting of ownership sector, industry group, and occupation group.
Source: Authors' calculations using data from the Office of Personnel Management.

Table 3: Wage index calculations of full-time and part-time work schedules, 2020 second quarter to 2022 second quarter

Work schedule	**Period**	**Laspeyres**	**Paasche**	**Fisher**
Full time	2020 Q2 to 2021 Q2	1.0130	1.0129	1.0129
	2021 Q2 to 2022 Q2	1.0337	1.0337	1.0337
Part time	2020 Q2 to 2021 Q2	1.0366	1.0361	1.0363
	2021 Q2 to 2022 Q2	1.0427	1.0425	1.0426

Note: Q2 = second quarter.
Source: Authors' calculations using data from the Office of Personnel Management.

Table 4: Wage index calculations, by Census divisions, 2020 second quarter to 2022 second quarter

Census division	**Period**	**Laspeyres**	**Paasche**	**Fisher**
New England	2020 Q2 to 2021 Q2	1.0123	1.0122	1.0122
	2021 Q2 to 2022 Q2	1.0491	1.0490	1.0490
Middle Atlantic	2020 Q2 to 2021 Q2	1.0144	1.0144	1.0144
	2021 Q2 to 2022 Q2	1.0355	1.0355	1.0355
East South Central	2020 Q2 to 2021 Q2	1.0183	1.0184	1.0184
	2021 Q2 to 2022 Q2	1.0370	1.0368	1.0369
South Atlantic	2020 Q2 to 2021 Q2	1.0097	1.0097	1.0097
	2021 Q2 to 2022 Q2	1.0320	1.0320	1.0320
East North Central	2020 Q2 to 2021 Q2	1.0159	1.0158	1.0159
	2021 Q2 to 2022 Q2	1.0341	1.0341	1.0341
West North Central	2020 Q2 to 2021 Q2	1.0152	1.0152	1.0152
	2021 Q2 to 2022 Q2	1.0378	1.0375	1.0377
West South Central	2020 Q2 to 2021 Q2	1.0145	1.0145	1.0145
	2021 Q2 to 2022 Q2	1.0345	1.0346	1.0345
Mountain	2020 Q2 to 2021 Q2	1.0150	1.0150	1.0150
	2021 Q2 to 2022 Q2	1.0398	1.0395	1.0397
Pacific	2020 Q2 to 2021 Q2	1.0207	1.0206	1.0206
	2021 Q2 to 2022 Q2	1.0437	1.0436	1.0437

Note: Q2 = second quarter.
Source: Authors' calculations using data from the Office of Personnel Management.

Table 5: Wage index calculations, by Census region, 2020 second quarter to 2022 second quarter

Census region	Period	Laspeyres	Paasche	Fisher
Northeast	2020 Q2 to 2021 Q2	1.0139	1.0139	1.0139
	2021 Q2 to 2022 Q2	1.0388	1.0388	1.0388
South	2020 Q2 to 2021 Q2	1.0155	1.0155	1.0155
	2021 Q2 to 2022 Q2	1.0361	1.0360	1.0360
Midwest	2020 Q2 to 2021 Q2	1.0111	1.0111	1.0111
	2021 Q2 to 2022 Q2	1.0322	1.0322	1.0322
West	2020 Q2 to 2021 Q2	1.0180	1.0179	1.0180
	2021 Q2 to 2022 Q2	1.0412	1.0411	1.0412

Note: Q2 = second quarter.
Source: Authors' calculations using data from the Office of Personnel Management.

Table 6: Wage index calculations by size class, 2020 second quarter to 2022 second quarter

Size class	Period	Laspeyres	Paasche	Fisher
1 (< 50)	2020 Q2 to 2021 Q2	1.0189	1.0189	1.0189
	2021 Q2 to 2022 Q2	1.0402	1.0402	1.0402
2 (51 to 100)	2020 Q2 to 2021 Q2	1.0117	1.0116	1.0116
	2021 Q2 to 2022 Q2	1.0407	1.0408	1.0408
3 (101 to 500)	2020 Q2 to 2021 Q2	1.0104	1.0104	1.0104
	2021 Q2 to 2022 Q2	1.0349	1.0349	1.0349
4 (> 500)	2020 Q2 to 2021 Q2	1.0133	1.0132	1.0132
	2021 Q2 to 2022 Q2	1.0342	1.0341	1.0342

Note: Q2 = second quarter.
Source: Authors' calculations using data from the Office of Personnel Management.

Table 7: Comparison of federal Laspeyres index with official Employer Cost Index, 2020 second quarter to 2022 second quarter

Period	Private industry	State and local	Exploratory federal
2020 Q2 to 2021 Q2	1.0315	1.0202	1.0131
2021 Q2 to 2022 Q2	1.0554	1.0341	1.0342

Note: Q2 = second quarter.
Source: Authors' calculations using data from the Office of Personnel Management.

Conclusions

This analysis demonstrates the practicality of using OPM data to compute a federal government wage component of the ECI. Other elements of the ECI may also be feasible if benefit-cost and hours data can be acquired. Given the magnitude of the U.S. federal workforce, its inclusion would expand NCS coverage as well as filling a void in information about federal workers. Although the annually announced federal pay increase provides some information about federal employment cost growth, it is an imprecise indicator—actual cost growth depends on the flow of employees into and out of federal service and the mix of employee tenures. The calculation of a wage or employment cost index would provide BLS data users useful measures of the growth of federal employment costs.

Further exploration of OPM data for use with the NCS will be enhanced by access to benefit-cost data. Even though acquiring benefit-cost data might be infeasible, we believe that the construction of federal wage indexes would prove a valuable addition to the NCS. The addition of the federal workforce to the NCS will require an analyst-validated NAICS crosswalk, which we view to be an attainable goal considering the findings presented in this article.

Appendix: North American Industry Classification System codes by industry and Standard Occupational Classification codes by occupation

Table A-1: Government industry group definitions, including codes

PNAICS	NAICS	Industry
G000	21, 23, 31-33	Goods Producing
4400	221	Utilities
420A	42-45	Wholesale & Retail Trade
4300	48,49	Transportation and warehousing
6110	6111	Elementary and secondary schools
6112	6112	Junior colleges
6113	6113	Colleges, universities, and professional schools
61R0	61, excl. 6111-6113	Rest of educational services
6220	622	Hospitals
6230	623	Nursing and residential care facilities
62R0	621, 624	Rest of Health Services
9200	92, excl. 928	Public administration
81R0	51-56, 71-81 excl. 814	Rest of Services

Note: PNAICS = pseudo-North American Industry Classification System, and NAICS = North American Industry Classification System.
Source: U.S. Bureau of Labor Statistics.

Table A-2: Occupation group definitions, including codes

PSOC	SOC	Occupation
110	11, 13	Management, business, and financial
120	15, 17, 19, 21, 23, 25, 27, 29	Professional and related
210	41	Sales and related
220	43	Office and administrative support
300	31-39	Service
405	45, 47	Farm, Fishing, Forestry, Construction, and Extraction
430	49	Installation, maintenance and repair
510	51	Production
520	53	Transportation and material moving

Note: PSOC = pseudo-Standard Occupational Classification, and SOC = Standard Occupational Classification.
Source: U.S. Bureau of Labor Statistics.

Notes

[1]Each basic Employer Cost Index (ECI) "cell" is categorized into industry and occupation groups. ECI cells are further separated into subcategories or "subcells." These subcategories include full- or part-time work, Census division or region, establishment size, metropolitan or nonmetropolitan, New York–Chicago–Los Angeles area, union status, and time and incentive status. Our analysis includes only subcells for full- or part-time work, Census division, region, and establishment size.

[2]An establishment is defined as an economic unit that produces goods or services, usually at a single physical location, and that is engaged in one or predominantly one type of economic activity. For more information, see U.S. Bureau of Labor Statistics glossary.

[3]U.S. Census, "General information about price indexes" (U.S. Census Bureau, n.d.).

[4]For a list of index formulas, see Wikipedia: The Free Encyclopedia, "List of price index formulas,"; and U.S. Census, "General information about price indexes."

[5]Typically, Laspeyres and Paasche index number formulas are expressed as a ratio of total wage costs, given period-0 and period-1 fixed employment baskets

$$I_L = \frac{\sum_{i=1}^{n} w_i^1 e_i^0}{\sum_{j=1}^{n} w_j^0 e_j^0} \tag{5}$$

and

$$I_P = \frac{\sum_{i=1}^{n} w_i^1 e_i^1}{\sum_{j=1}^{n} w_j^0 e_j^1}. \tag{6}$$

After some manipulation of these formulas, the Laspeyres and Paasche indexes can also be expressed as the function of wage relatives and expenditure shares, as given in the main text.

[6] Since the Törnqvist and Fisher indexes are close approximations of one another (formulas produce numbers that are close to one another), we do not use the slightly more complicated Törnqvist index number formula.

[7]OPM defines part-time work as between 16 and 32 hours a week and full-time work as more than 32 hours a week. In addition to full-time and part-time work, a number of other work schedules include full-time seasonal, part-time seasonal, intermittent, and intermittent seasonal. Our analysis only includes full-time and part-time workers.

[8]We excluded military bases because they can have establishments such as schools, hospitals, entertainment venues, and so forth. Although nurses and teachers might be straightforward to classify into hospitals and schools, occupations such as janitors and secretaries would be challenging. U.S. Postal Service employee data are separately available from OPM and potentially could be included in the future.

[9] "Fact sheet: computing hourly rates of pay using the 2,087-hour divisor" (U.S. Office of Personnel Management, n.d.).

[10]Underlying the ECI is the Laspeyres index number formula.

[11]Technically, ECI jobs are also differentiated by union status and time

or incentive status. Union status is unavailable in our data, and to our knowledge, incentive pay is not widely used in the federal government.

[12]Michael K. Lettau, Mark A. Loewenstein, and Aaron Cushner, "Is the ECI sensitive to the method of aggregation?" *Monthly Labor Review*, June 1997; and Michael K. Lettau, Mark A. Loewenstein, and Steve P. Paben, "Is the ECI sensitive to the method of aggregation? an update," *Monthly Labor Review*, December 2002.

[13]Ibid. For an explanation of this pattern reversal, see specifically Lettau et al. "Is the ECI sensitive to the method of aggregation?"