The Evolving U.S. Occupational Structure

Enghin Atalay[1] Phai Phongthiengtham[1]
Sebastian Sotelo[2] Daniel Tannenbaum[3]

[1]University of Wisconsin-Madison, [2]University of Michigan-Ann Arbor,

[3]University of Nebraska-Lincoln

September, 2017

Motivation

- Labor income inequality has increased over the last several decades.
  - 90-10 ratio of real earnings $\uparrow$27 log points between 1960 and 2000.
- The task approach to labor market:
  - Better, cheaper computers
  - Easier offshoring
  - $\Longrightarrow$Changes in the demand for particular (routine, offshorable) tasks
- To assess the task-based model, past work combines:
  - occupations' task content, measured at given point in time (e.g., O*NET)
  - shifts in employment shares across occupations
- Occupations rich in routine tasks have shrunk, those centered on non-routine interactive tasks have grown

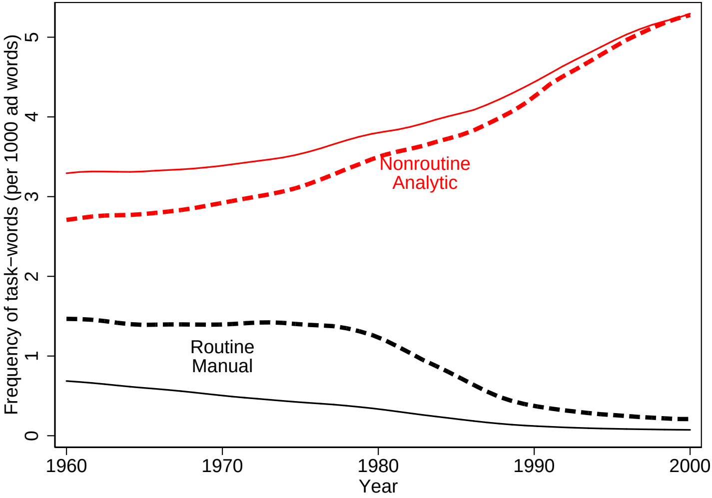
Research Questions

- Are there within-occupations trends in the tasks which workers perform?
- Can within-occupation changes in task content help explain increasing earnings inequality?

What we do: Data

- Measure evolution of within-occupation task content using the frequency of task-related words in job ads
- Construct a new data set drawing from the text of newspaper vacancy postings:
  - $\sim$ 4.1 million ads
  - 1960-2000
  - New York Times, Wall Street Journal and Boston Globe

Decline of routine manual tasks, rise of nonroutine analytic



Notes. Dashed lines: Non-supervisory production workers. Solid lines: All workers

What we do: Main results

- Consistent with previous research:
  - Routine tasks have declined markedly while nonroutine tasks have become increasingly more important
- Large share of aggregate changes in task content occurred within occupations, rather than between.
- Use changes in the task composition of occupations to account for inequality between 1960 and 2000
  - Using equilibrium methods: account for 22 log point increase in 90-10 earnings inequality
  - Using statistical decomposition: account for changes of a similar magnitude

Literature

- **Task approach**: Autor, Levy, Murnane (2003), Autor and Dorn (2013), Acemoglu and Autor (2011), Goos, Manning, Salomons (2014), Spitz-Oener (2006), Firpo, Fortin, Lemieux (2014), Becker and Muendler (2015), and many other recent studies.
  - Contribution: Time-varying task measurements
- **On-line vacancy postings**: Deming and Kahn (2016), Hershbein and Kahn (2016), Marinescu and Wolthoff (2016), Modestino, Shoag, Ballance (2016).
  - Contribution: Long-run measurements (pre-internet era)
- **Comparative advantage and occupational choice**: Heckman and Sedlaceck (1985), Heckman and Scheinkman (1987), Burstein, Morales, Vogel (2015), and Hsieh, Hurst, Jones, Klenow (2016)
  - Contribution: embed task bundling in a quantitative, GE model of sorting

Roadmap

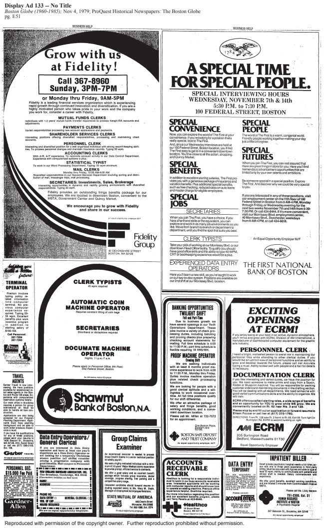
1. Turning job ads into data
2. Trends in task-related words
3. Tasks and the earnings distribution

Processing newspaper text files

- ProQuest processes images of newspaper pages into text files (OCR)
- Job ads from New York Times (1960-2000), Wall Street Journal (1960-1998), and Boston Globe (1960-1983)
- Steps to construct the data set:

1. Distinguish vacancy postings from other advertisements
2. Find the boundaries between vacancy postings
3. Identify the ad's job title $\Rightarrow$SOC code
4. Extract task-related information

Processing newspaper text files - Unprocessed


Display Ad 133 -- No Title
Boston Globe (1960-1985); Nov 4, 1979; ProQuest Historical Newspapers: The Boston Globe
pg. E51

BUSINESS HELP

Grow with us at Fidelity!

Call 367-8960
Sunday, 3PM-7PM
or Monday thru Friday, 9AM-5PM

MUTUAL FUNDS CLERKS

PAYMENTS CLERKS

SHAREHOLDER SERVICES CLERKS

PERSONNEL CLERK

ACCOUNTING CLERKS

STATISTICAL TYPIST

MAIL CLERKS

SECRETARIES: Investments, Sales, Brokerage

We encourage you to grow with Fidelity and share in our success.

Fidelity Group
82 DEVONSHIRE STREET
BOSTON, MA 02109

A SPECIAL TIME FOR SPECIAL PEOPLE.

SPECIAL INTERVIEWING HOURS
WEDNESDAY, NOVEMBER 7th & 14th
5:30 P.M. to 7:30 P.M.
100 FEDERAL STREET, BOSTON

SPECIAL CONVENIENCE

SPECIAL PEOPLE

SPECIAL FUTURES

SPECIAL BENEFITS

SPECIAL JOBS

SECRETARIES

CLERK TYPISTS

EXPERIENCED DATA ENTRY OPERATORS

An Equal Opportunity Employer M/F

THE FIRST NATIONAL BANK OF BOSTON

TERMINAL OPERATOR

CLERK TYPISTS

AUTOMATIC COIN MACHINE OPERATOR

SECRETARIES

DOCUMATE MACHINE OPERATOR

Shawmut Bank of Boston, N.A.

TRAVEL AGENTS

Garber

BANKING OPPORTUNITIES
TWILIGHT SHIFT

PROOF MACHINE OPERATOR

BOSTON SAFE DEPOSIT AND TRUST COMPANY

EXCITING OPENINGS AT ECRM!

PERSONNEL CLERK

DOCUMENTATION CLERK

ECRM
205 Burlington Road
Bedford, Massachusetts 01730
Equal Opportunity Employer

PERSONNEL SEC.
$15,600 Fee Paid

Gardner-Allen

Data Entry Operators/ General Clerical

Group Claims Examiner

STATE MUTUAL OF AMERICA

ACCOUNTS RECEIVABLE CLERK

Healthco

DATA ENTRY TEMPORARY

accountemps

INPATIENT BILLER

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Note: From the Boston Globe, November 4, 1979

Processing newspaper text files – Unprocessed

```
rapid growth through continued innovation and
diversification If you are highly motivated person who takes
pride in your work and the company you work for consider
career with Fidelity \n MUTUAL FUNDS CLERKS \n
Individuals with 1-2 years funds transfer experience to
process Keogh IRA accounts and adjustments \n PAYMENTS
CLERKS \n Varied les processing new account applications and
payments \n StI nt flt ner fr nlr ct tfl \n Mnirk \n and
maintaining client \n strong record-keeping 50 wpm \n Data
Control Department \n sorting and \n Brokerage \n
environments with \n benefits package for our Boston
convenient to the Market \n gr WilnliE lU5lty \n success \n
-l 1?Q1.a1 ol P-1Sl1v MPloV \n Fidelit \n Group \n
82 DEVONSHIRE STREET BOSTON MA 02109 \n 111 \n -l \n
TERMINAL OPERATOR \n
Position involves typing policy related information Into
computer terminal No previous computer experience required
Typing 5055 wpm Excellent benefits plus work Incentive
program in addition to starting salary of S150-165.
```

Note: Snippet of the raw text from Boston Globe, 11/4/79, Display Ad #133

Processing newspaper text files – Processed

```
rapid growth through continued innovation and diversification If you are
highly motivated person who takes pride in your work and the company you
work for consider career with Fidelity \n
--------------------------------------------------------------
MUTUAL FUNDS CLERKS \n
Individuals with 1-2 years funds transfer experience to process Keogh IRA
accounts and adjustments \n
--------------------------------------------------------------
PAYMENTS CLERKS \n
Varied les processing new account applications and payments \n StI nt flt
ner fr nlr ct tfl \n Mnirk \n and maintaining client \n strong record-
keeping 50 wpm \n Data Control Department \n sorting and \n Brokerage \n
environments with \n benefits package for our Boston convenient to the
Market \n gr WilnliE lU5lty \n success \n -l 1?Q1.a1 ol P-1Sl1v MPloV \n
Fidelit \n Group \n
82 DEVONSHIRE STREET BOSTON MA 02109 \n
--------------------------------------------------------------
111 \n -l \n
--------------------------------------------------------------
TERMINAL OPERATOR \n
Position involves typing policy related information Into computer
terminal No previous computer experience required Typing 5055 wpm
Excellent benefits plus work Incentive program in addition to
starting salary of S150-165.
```

Note: Snippet of the raw text from Boston Globe, 11/4/79, Display Ad #133

Most common occupations in newspaper ads

Job Title		4-Digit SOC Occupations	
Description	Count	Description	Count
Secretary	117.3	4360: Secretary	408.6
Sales	55.4	4390: Oth. Admin. Support	340.3
Assistant	53.6	1320: Accountant	194.6
Accounting	51.0	1511: Computer Sci.	185.4
Clerk	50.0	4330: Financial Clerks	174.8
Accounting	50.4	4330: Bookkeeper	159.5
Engineer	42.8	2911: Nurse	124.8
Manager	41.8	1110: General Mgr.	120.2
Bookkeeper	41.3	1721: Engineer	119.1
Salesperson	40.2	1130: Financial Mgr.	90.5

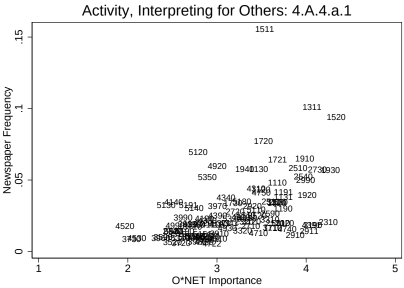
Notes: This table lists the top five job titles (columns 1-2), and the top 10 4-digit SOC codes (columns 3-4) in our Boston Globe, New York Times, and Wall Street Journal data. The counts are given in thousands of newspaper job ads.

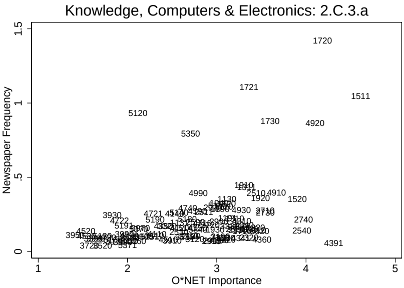
Four mappings of words to task classifications

We group tasks and skills according to four alternative classifications:

1. Spitz-Oener (2006)
   - 5 categories: 1) non-routine analytic, 2) non-routine interactive, 3) non-routine manual, 4) routine cognitive and 5) routine manual
2. O*NET
   - 16 Work Styles (e.g., Achievement/Effort, Adaptability/Flexibility)
   - 35 Skills (Active Learning, Active Listening, Complex Problem Solving)
   - 33 Knowledge Requirements (Administration and Management, Biology)
   - 41 Activities (Assisting and Caring for Others)
3. Firpo, Fortin, and Lemieux (2014)
   - 4 categories: information content, face-to-face contact, on-site job, decision-making
4. Deming and Kahn (2016)

Comparison to O*NET





corr. $= 0.51$ corr. $= 0.54$

- Correlation with O*NET elements mostly in the 0.4 to 0.7 range (across occupations)
- Non-linear relationship: Rarely mentioned keywords for O*NET low importance values.

Mapping to Spitz-Oener (2006) classification

non-routine analytic	analyze, analyzing, design, designing, devising rule, evaluate, evaluating, interpreting rule, plan, planning, research, researching, sketch, sketching
non-routine interactive	advertise, advertising, advise, advising, buying, coordinate, coordinating, entertain, entertaining, lobby, lobbying, managing, negotiate, negotiating, organize, organizing, presentation, presentations, presenting, purchase, sell, selling, teaching
non-routine manual	accommodate, accommodating, accommodation, renovate, renovating, repair, repairing, restore, restoring, service, serving
routine cognitive	bookkeeping, calculate, calculating, correcting, corrections, measurement, measuring
routine manual	control, controlling, equip, equipment, equipping, operate, operating

*we include for each of these words, synonyms based on machine-learning text similarity

Ranking occupations by task content

Nonroutine Analytic			Nonroutine Interactive		
1720:	Engineers	0.89	1120:	Sales Managers	1.11
1721:	Engineers	0.89	4140:	Sales Rep., Whole./Man.	0.80
1930:	Social Scientists	0.70	4130:	Sales Rep., Services	0.64
Nonroutine Manual			**Routine Cognitive**		
4930:	Vehicle Mechanics	0.38	4330:	Financial Clerks	0.22
4910:	Maint. Supervisors	0.29	4390:	Other Admin. Support	0.10
4990:	Other Maintenance	0.29	4320:	Comm. Equip. Operators	0.09
Routine Manual					
5140:	Metal and Plastic	0.14			
4990:	Other Maintenance	0.07			
5141:	Metal and Plastic	0.06			

Notes: (1) Occupation code, (2) Occupation title, (3) Mentions per ad

Relative importance of within versus between occupations

- Question: What fraction of changes in aggregate use of task h are driven by changes in tasks within occupations?
- Aggregate use of task h at time t is $\bar{T}_{ht} = \sum_j \vartheta_{jt} \widetilde{T}_{hjt}$
- We decompose aggregate changes:

$$\bar{T}_{ht} - \bar{T}_{h,1960} = \underbrace{\sum_j \vartheta_{j,1960} \left(\widetilde{T}_{hjt} - \widetilde{T}_{hj,1960} \right)}_{\text{within}} + \underbrace{\sum_j \left(\vartheta_{jt} - \vartheta_{j,1960} \right) \widetilde{T}_{hjt}}_{\text{between}}$$

where

- $\tilde{T}_{hjt}$ is mentions of task h per thousand ad words for occupation j in year t.
- ϑ_{jt} share of employment in occupation j at time t (Census data)

Trends in keyword frequencies (Spitz-Oener, 2006 class.)

	Frequency					
	60-64	65-69	75-79	80-84	90-94	95-00
Nonroutine Analytic	3.19	3.50	3.63	3.93	4.63	5.48
Nonroutine Interactive	4.94	4.56	5.18	5.84	6.56	7.03
Nonroutine Manual	2.50	2.36	2.54	2.59	2.59	2.73
Routine Cognitive	1.00	0.91	0.65	0.74	0.59	0.54
Routine Manual	0.70	0.58	0.42	0.27	0.10	0.07

Trends in keyword frequencies (Spitz-Oener, 2006 class.)

	Frequency						Within
	60-64	65-69	75-79	80-84	90-94	95-00	Share
Nonroutine	3.19	3.50	3.63	3.93	4.63	5.48	0.89
Analytic	[3.16]	[3.47]	[3.49]	[3.71]	[4.32]	[5.22]	(0.04)
Nonroutine	4.94	4.56	5.18	5.84	6.56	7.03	0.94
Interactive	[4.94]	[4.52]	[5.07]	[5.85]	[6.29]	[6.89]	(0.05)
Nonroutine	2.50	2.36	2.54	2.59	2.59	2.73	1.49
Manual	[2.52]	[2.40]	[2.56]	[2.65]	[2.54]	[2.86]	(0.68)
Routine	1.00	0.91	0.65	0.74	0.59	0.54	0.98
Cognitive	[0.99]	[0.90]	[0.61]	[0.73]	[0.61]	[0.54]	(0.04)
Routine	0.70	0.58	0.42	0.27	0.10	0.07	0.97
Manual	[0.71]	[0.60]	[0.50]	[0.31]	[0.11]	[0.10]	(0.04)

The relation between task changes and inequality

1. Quantitative GE model of sorting and comparative advantage
2. Decompositions via RIF regressions

Quantitative GE model – Environment

- Time t
- Task space $h = 1, \ldots, H$
- Worker groups $g = 1, \ldots, G$:
  - mass L_{gt}, each endowed with one unit of labor
  - skills S_{gh}
- Occupations, $j = 1, \ldots, J$

Quantitative GE model – Technology

- Worker i from group g produces task h according to

$$q_{ight} = S_{gh} l_{ight}$$

with l_{ight} time allocated to task h

- Occupation production function

$$V_{igjt} = \epsilon_{igjt} \prod_{h} \left(\frac{q_{ight}}{T_{hjt}} \right)^{T_{hjt}}$$

where ϵ_{igjt} is an idiosyncratic efficiency shock

Quantitative GE model – Preferences

- Consumers combine occupation output according to

$$U = \left(\sum_{j=1}^{J} \sigma_j^{1/\rho} C_j^{\frac{\rho-1}{\rho}} \right)^{\frac{\rho}{\rho-1}}$$

Quantitative GE model – Equilibrium

- Workers are rewarded with all their value added

- Conditional on choosing occupation j, allocation of time

$$\max_{\{l_{ight}\}} P_{jt} V_{igjt} \left(\{l_{ight}\}\right)$$

subject to

$$\sum_h l_{ight} = 1$$

- Optimal time allocation

$$l_{ight} = T_{hjt} \cdot \left(\sum_{h'} T_{h'jt} \right)^{-1}$$

same for all i, g

Quantitative GE model – Equilibrium

- Wage equation

$$\log W_{igjt} = \log P_{jt} + \sum_{h=1}^{H} T_{hjt} \log S_{gh} + \log \epsilon_{igjt}$$

- Frechet unobserved ability: $\Pr\left[\epsilon_{igjt} < z\right] = \exp\left(-z^{-\theta}\right)$

- Optimal sorting

$$\lambda_{gjt} = \phi_{gjt} / \sum_{j'} \phi_{gj't}$$

$$\phi_{gjt} = P_{jt}^{\theta} \prod_{h=1}^{H} S_{gh}^{\theta T_{hjt}}$$

- Average wages per group

$$\overline{W}_{gt} = \Gamma\left(1 - \frac{1}{\theta}\right) \cdot \left(\sum_{j'} \phi_{gjt}\right)^{1/\theta}$$

Quantitative GE model – Equilibrium

- Equilibrium at time t: occupational prices P_{jt}, such that market clears for each occupation j

$$\frac{\sigma_j \cdot (P_{jt})^{1-\rho}}{\sum_{j'} \sigma_{j'} \left(P_{j't}\right)^{1-\rho}} \sum_g \overline{W}_{gt} L_{gt} = \sum_{g=1}^{G} \overline{W}_{gt} \lambda_{gjt} L_{gt}$$

recall

$$\overline{W}_{gt} = \Gamma\left(1 - \frac{1}{\theta}\right) \cdot \left(\sum_{j'} \left(P_{j't} \prod_{h=1}^{H} (S_{gh})^{T_{hj't}}\right)^{\theta}\right)^{1/\theta}$$

Quantitative GE model – Estimation

- Parameterize task skills

$$\log S_{gh} = a_{h,gender} D_{gender,g} + a_{h,edu} D_{edu,g} + a_{h,exp} D_{exp,g}$$

- Method of moments estimation. Objective:

$$\sum_g \sum_j \omega_{gj}^{\lambda} \left(\log \lambda_{gj,1960} - \log \lambda_{gj,1960}^{data} \right)^2 + \sum_g \omega_g^W \left(\log \bar{W}_{g,1960} - \log \bar{W}_{g,1960}^{data} \right)^2$$

  using data for $t = 1960$, ω_{gj}^{λ} and ω_g^W inversely related to variance

- We use data on $G \times (J-1) + G = 40 \times 81 = 3240$ moments to identify J fixed effects and 40 a parameters

- Fix $\theta = \rho = 1.78$ (Burstein et al. 2017)

Quantitative GE model – Estimation

	Nonroutine Analytic	Nonroutine Interactive	Nonroutine Manual	Routine Cognitive	Routine Manual
Gender					
Female	-1.384	-0.473	0.269	3.472	-7.420
	(0.004)	(0.004)	(0.005)	(0.006)	(0.017)
Education					
$<$ HS	-1.981	-1.155	2.628	-2.199	4.461
	(0.007)	(0.005)	(0.011)	(0.012)	(0.034)
High School	-0.785	-0.498	0.905	-0.006	2.757
	(0.008)	(0.005)	(0.011)	(0.010)	(0.033)
College	1.449	0.106	1.414	-1.002	-8.438
	(0.008)	(0.006)	(0.017)	(0.016)	(0.057)
Post-Graduate	1.394	-0.215	3.222	-2.778	-7.300
	(0.008)	(0.007)	(0.011)	(0.020)	(0.059)
Experience					
0-9 Years	-0.068	-0.602	-0.365	0.586	-2.560
	(0.005)	(0.005)	(0.007)	(0.010)	(0.020)
10-19 Years	-0.002	-0.219	0.216	0.354	-0.906
	(0.006)	(0.005)	(0.007)	(0.012)	(0.018)
30+ Years	-0.063	0.153	-0.026	0.239	-0.767
	(0.006)	(0.004)	(0.007)	(0.011)	(0.020)

Notes: To compute the standard errors, we re-sampled 40 times from the 1960 decennial census.
Fit: Correlation of log employment shares: .56; correlation of log wages: .95

Quantitative GE model – Estimation

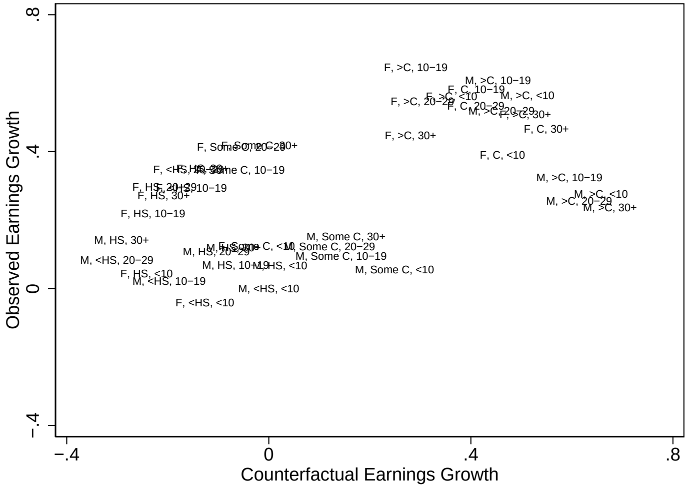
	Nonroutine Analytic	Nonroutine Interactive	Nonroutine Manual	Routine Cognitive	Routine Manual
Gender					
Female	-1.384	-0.473	0.269	3.472	-7.420
	(0.004)	(0.004)	(0.005)	(0.006)	(0.017)
Education					
$<$ HS	-1.981	-1.155	2.628	-2.199	4.461
	(0.007)	(0.005)	(0.011)	(0.012)	(0.034)
High School	-0.785	-0.498	0.905	-0.006	2.757
	(0.008)	(0.005)	(0.011)	(0.010)	(0.033)
College	1.449	0.106	1.414	-1.002	-8.438
	(0.008)	(0.006)	(0.017)	(0.016)	(0.057)
Post-Graduate	1.394	-0.215	3.222	-2.778	-7.300
	(0.008)	(0.007)	(0.011)	(0.020)	(0.059)
Experience					
0-9 Years	-0.068	-0.602	-0.365	0.586	-2.560
	(0.005)	(0.005)	(0.007)	(0.010)	(0.020)
10-19 Years	-0.002	-0.219	0.216	0.354	-0.906
	(0.006)	(0.005)	(0.007)	(0.012)	(0.018)
30+ Years	-0.063	0.153	-0.026	0.239	-0.767
	(0.006)	(0.004)	(0.007)	(0.011)	(0.020)

Notes: To compute the standard errors, we re-sampled 40 times from the 1960 decennial census.
Fit: Correlation of log employment shares: .56; correlation of log wages: .95

Quantitative GE model – Counterfactual

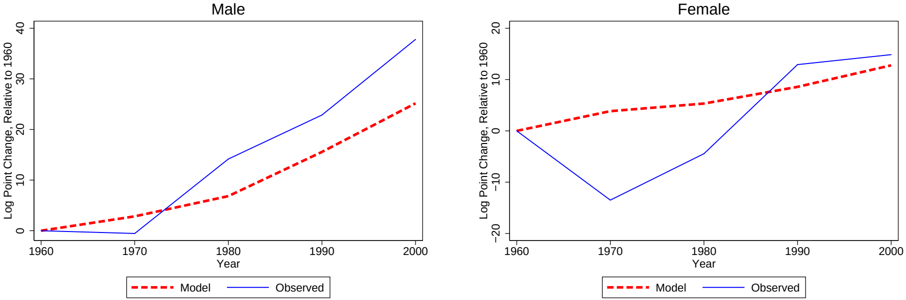
- Q: What would happen if changes in task demand were exogenous and the only thing that happened?
- Introduce changes in T_{hjt}, between 1960 and 2000 and measure change in the wage distribution

Quantitative GE model – Counterfactual



Notes: Each point gives the growth in wages for one of the 40 g groups The first character—"M" or "F"—describes the gender; the second set of characters—"<HS," "HS," "Some C," "C," or ">C"—the educational attainment; and the third set of characters the number of years of potential experience for the demographic group.

Quantitative GE model – Counterfactual



Notes: Each panel plots the change in 90-10 earnings inequality, as observed in the decennial census (solid line), and due only to changes in the demand for tasks (dashed line).

Quantitative GE model - Discussion

- Advantages
  - Allows us to isolate the effects of shocks to task demands
  - Account for all GE adjustments
  - Acknowledge that optimal sorting has an effect on inequality
- Disadvantages
  - Only generates wage variation between groups
  - Conditional on P_j, task pricing does not contribute to wage inequality
  - Take changes in T as exogenous relative to other forces in the economy (labor supply, shifts in output demand, etc.)

Tackling the problem with decomposition methods

- Q: Can we account for changes in the earnings distribution,
  - keeping implicit task prices fixed, but allowing for changes in worker and job characteristics?
  - keeping the distribution of characteristics fixed, but allowing for changes in task prices?
- Follow method in Fortin, Firpo, and Lemieux (2011)
  - Decompose change in the distributional statistic ν, Δ_O^{ν}
  - If ν is the mean, FFL decomposition is the Oaxaca-Blinder decomposition:

$$\Delta_O^{\nu} = \bar{w}_1 - \bar{w}_0 = \underbrace{\left[\sum_{k=1}^{K} \bar{X}_{1k}\left(\beta_{1k} - \beta_{0k}\right)\right]}_{\text{wage structure effect}} + \underbrace{\left[\sum_{k=1}^{K} \left(\bar{X}_{1k} - \bar{X}_{0k}\right)\beta_{0k}\right]}_{\text{composition effect}}$$

  - Covariates: race, education, experience and task contents (routine vs non-routine tasks) at 4 digit SOC level

FFL decomposition as a generalization of Oaxaca-Blinder

- We perform FFL decomposition for $\nu = \tau$th quantile
- Replace outcome with the Re-centered Influence Function (RIF) of ν

$$\mathsf{RIF}(w;\tau) \equiv w_\tau + (\tau - \mathbf{1}(\mathrm{w} < \mathrm{w}_\tau)) / f(w_\tau)$$

which satisfies $\mathbb{E}[\mathsf{RIF}(w;\tau)] = w_\tau$.

- Note

$$\begin{aligned} \mathbb{E}[\mathsf{RIF}(w;\tau)|X] &= c_{1\tau}\mathbb{E}[1(\mathrm{w} < \mathrm{w}_\tau)|X] + c_{2\tau} \\ &= c_{1\tau}\Pr[w < w_\tau|X] + c_{2\tau} \end{aligned}$$

- **Assumption:**

$$\mathbb{E}[\mathsf{RIF}(w;\tau)|X] = X\gamma^\tau$$

FFL decomposition is a generalization of Oaxaca-Blinder

Proceed in steps for each decile τ

1. Construct the RIF for τ
2. Obtain $\hat{\gamma}_t^\tau$ from

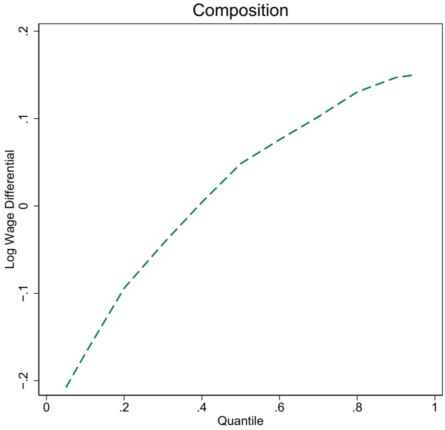
$$\mathbb{E}\left[\mathrm{RIF}\left(w;\tau\right)|X\right] \quad = \quad X_t\gamma_t^\tau$$

   - Regression coefficient of $1\left(\mathrm{w} < \mathrm{w}_\tau\right)$ on X: change in the earnings quantile from a one unit change in X.
   - $f\left(w_\tau\right)$: change in the earnings quantile from a one unit change in earnings.
   - $\hat{\gamma}_t^\tau$: change in the unconditional earnings distribution at τth quantile from a unit increase in X.

3. Decompose the change in τ between $t=1$ and $t=0$ as

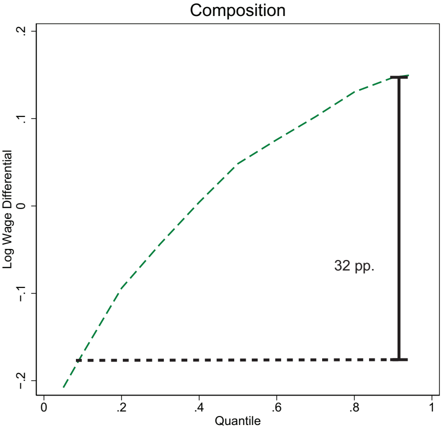
$$\Delta_O^\tau = w_1^\tau - w_0^\tau = \underbrace{\left[\sum_{k=1}^{K}\bar{X}_{1k}\left(\hat{\gamma}_{1k}^\tau - \hat{\gamma}_{0k}^\tau\right)\right]}_{\text{wage structure effect}} + \underbrace{\left[\sum_{k=1}^{K}\left(\bar{X}_{1k} - \bar{X}_{0k}\right)\hat{\gamma}_{0k}^\tau\right]}_{\text{composition effect}}$$

FFL Decomposition – Composition Effects from Task Measures



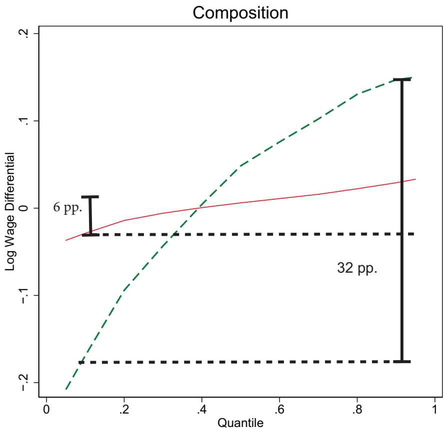
- Composition Effect From Task Measures: $\sum_{k \in \text{Spitz-Oener Tasks}} \left(T_{1k} - T_{0k}\right) \gamma_{0k}^{\tau}$

FFL Decomposition – Composition Effects from Task Measures



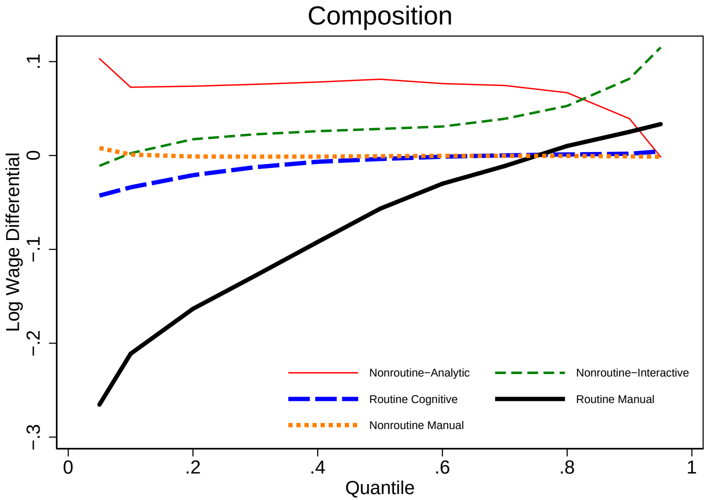
- Composition Effect From Task Measures: $\sum_{k \in \text{Spitz-Oener Tasks}} \left(T_{1k} - T_{0k}\right) \gamma_{0k}^{\tau}$

FFL Decomposition – Composition Effects from Task Measures

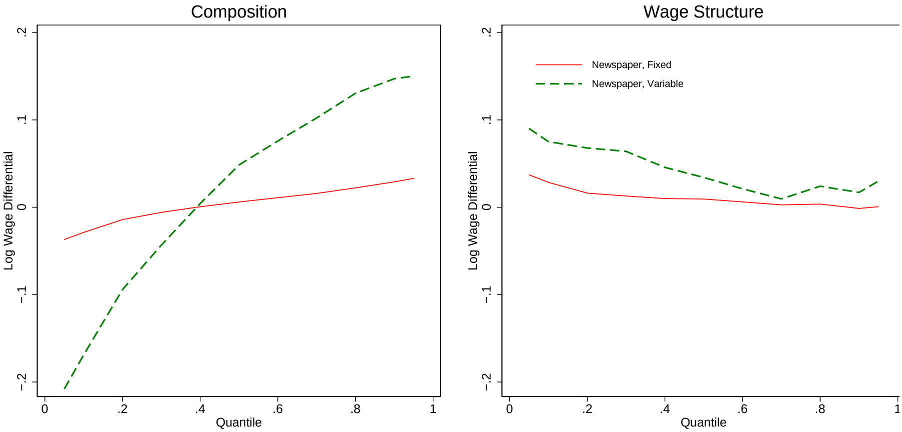


- Red line: Task measure for occupation j is fixed through the sample.

FFL Decomposition – Detailed Composition Effects



FFL Decomposition – Composition and Wage Structure Effects



FFL Decomposition - Discussion

- Advantages
  - Permits within-group wage changes resulting from task content shifts
- Disadvantages
  - Silent on GE adjustments.
  - Silent on sorting based on unobservable characteristics.

Conclusion

- New measurements of changes in task content of jobs over time
- "Within-occupation" changes are at least as important as "between-occupation" changes in accounting for aggregate changes in job content
- Reduced-form and model-based decompositions suggest our task measures account for about a 20 percentage point increase in the 90-10 ratio
- Future work: what drives these changes?
  - adoption of technology
  - specialization across cities

Appendix

Existing data sets

- DOTs: 5 editions (1939/1949/1965/1977/1991)
  - Not every edition contains numerical scores.
  - Status quo bias / vague, repetitive with ambiguous value-laden scales (Autor, 2013).
- O*NET
  - Complexity (400 rating scales) and loss of transparency (Autor, 2013).
  - Static
- German Qualification and Career Survey (IAB/BIBB)
  - Four cross sections: 1979, 1985/86, 1991/92, 1998/99
- Princeton Data Improvement Initiative (PDII): 2008

Distinguishing vacancy postings

- ProQuest has already provided a generic ad category in each newspaper block such as stock quote, obituary, editorial, display ads and classified ads.
- However, there are several types of advertisements within display ads and classified ads.
- We apply a Latent Dirichlet Allocation (LDA) algorithm.
  - There are several latent topics. Each topic has a different distribution of words.
  - The algorithm estimates these distributions and identifies which words appear more frequent in each topic.

Latent Dirichlet Allocation (LDA)

We proceed as follows:

1. Remove stop words (e.g., “a”, “is” and “and”), numerals and words that are not contained in the English dictionary.
2. Stem words (e.g., “management”, “managing”, “manages” and “manager” become “manag”)
3. Apply LDA and pick newspaper pages that has a probability of being job postings greater than 0.4.

Note that:

- We perform these steps separately for each of our newspapers: Display Ads/Classified Ads in Boston Globe/New York Times/Wall Street Journal.
- We found exceedingly few vacancy postings among the Wall Street Journal Display Ads and excluded them afterwards.
- We restored our original text afterwards.

Distinguishing vacancy postings

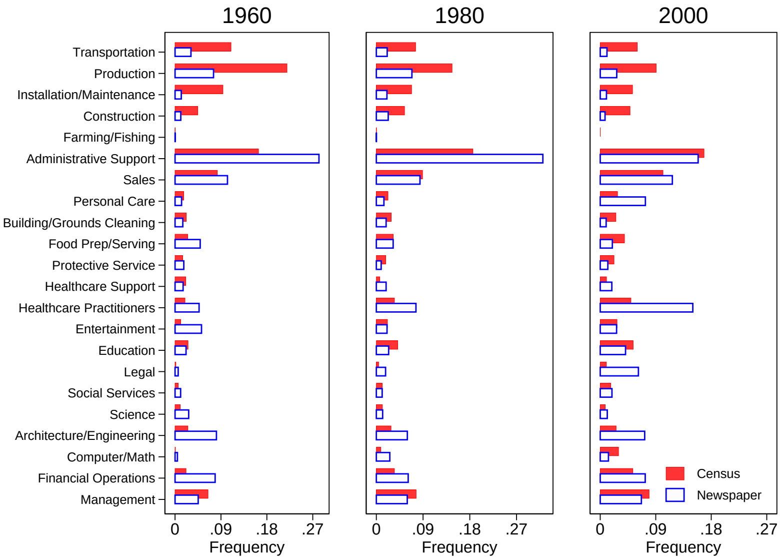
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	system	reg	car	day	street
2	experi	save	price	free	open
3	comput	store	new	new	inc
4	opportun	size	auto	one	ave
5	year	price	power	call	mass
6	manag	color	tire	coupon	call
7	engin	style	stereo	travel	rte
8	program	set	air	per	rout
9	requir	regular	door	offer	new
10	design	charg	stock	week	home

Note: This table provides the top 10 word stems for each of the 5 topics in our estimated LDA model on the Boston Globe Display Ad subsample.

Computing similarity among words/phrases

- Two problems to solve:
  - How to tell what job titles map to a particular SOC code?
  - How to tell what words in job ads map to a particular task?
- Use a continuous bag of words model to compute similarity among words/phrases in our newspaper text.
  - Idea: Words with similar meaning appear in similar contexts.
  - Example: "quantitative," "math" and "mathematical" are similar because all three appear close to "ability," "computation," "problem."

Newspaper vacancies are similar to Census employment shares



Newspaper vacancies and Census employment shares

- Relative to Census employment, our newspaper data set:
  - over-represents Sales, Health Practitioner, and Architecture, Engineering occupational groups
  - under-represents the Transportation, Production, and Installation and Maintenance occupational groups.
- Similar pattern as in on-line job postings (Hershbein and Kahn, 2016)
- We perform other checks on the data and argue selection of ads is not a major concern (e.g. checks on educational requirements)

Notation and Terminology

1. A vocabulary is a set of all possible V words.
2. A word w is a vector of length V. If w takes on the ith element of the vocabulary, then $w^i = 1$ and $w^j = 0$ for all $j \neq i$.
3. A document is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, ...w_N)$
4. A corpus is a collection of M documents denoted by $\mathbf{D} = \{\mathbf{w}_1, \mathbf{w}_2, ...\mathbf{w}_M\}$
5. A topic $z_k \in \{z_1, ..., z_K\}$ denotes a "hidden label" across documents in a corpus. The dimensionality K is assumed to be known and fixed.

Distinguishing vacancy postings

The model assumes generating process as follows.

1. First, a vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K)$ and a K by V matrix β are chosen and held fixed throughout the corpus.
2. Next, for each document $\mathbf{w}_m$ in the corpus, choose a K-dimensional topic weight vector $\theta_m \sim Dir(\alpha)$

$$p(\theta_{m1}, \theta_{m2}, ..., \theta_{mK} | \alpha_1, \alpha_2, ..., \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_i)}{\prod_{k=1}^{K} \Gamma(\alpha_i)} \prod_{k=1}^{K} \theta_{mk}^{\alpha_i - 1}$$

   where each $\alpha_i > 0$ and $\Gamma(x)$ is a gamma function.
3. Finally, each word w_{mn} in a document $\mathbf{w}_m$ is determined by:
   - Choose a topic $z_{mn} \in \{z_1, ..., z_K\}$ where $p(z_{mn} = z_k | \theta_{m1}, \theta_{m2}, ..., \theta_{mK}) = \theta_{mk}$.
   - Choose a word w_{mn} from word-topic probability matrix β where $\beta_{ij} = p(w^j = 1 | z^i = 1)$.

Distinguishing vacancy postings

Conditional on α and β, the joint distribution of a topic mixture θ_m, a set of topics $\mathbf{z}_m$ and a set of words $\mathbf{w}_m$ is given by:

$$p(\theta_m, \mathbf{z}_m, \mathbf{w}_m | \alpha, \beta) = p(\theta_m | \alpha) \prod_{n=1}^{N} p(z_{mn} | \theta_m) p(w_{mn} | z_{mn}, \beta)$$

The marginal distribution, or likelihood, of a document m is given by integrating over θ and summing over z:

$$\begin{aligned} p(\mathbf{w} | \alpha, \beta) &= \int p(\theta | \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n | \theta_m) p(w_n | z_n, \beta) \right) d\theta \\ &= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \right) \left(\prod_{n=1}^{N} \sum_{k=1}^{K} \prod_{j=1}^{V} (\theta_k \beta_{kv})^{w_m n} \right) d\theta \end{aligned}$$

Distinguishing vacancy postings

The main purpose of LDA is to determine the distribution of the latent topics conditional on the observed words in each document, which is:

$$p(\theta_m, \mathbf{z}_m | \mathbf{w}_m, \alpha, \beta) = \frac{p(\theta_m, \mathbf{z}_m, \mathbf{w}_m | \alpha, \beta)}{p(\mathbf{w}_m | \alpha, \beta)}$$

The estimated values $\hat{\alpha}$ and $\hat{\beta}$ are values of $\tilde{\alpha}$ and $\tilde{\beta}$ that maximize the log-likelihood of all the documents:

$$(\hat{\alpha}, \hat{\beta}) = \arg\max_{(\tilde{\alpha}, \tilde{\beta})} \left\{ \sum_{m=1}^{M} log \left[p(\mathbf{w}_m | \tilde{\alpha}, \tilde{\beta}) \right] \right\}$$

Discerning the boundaries between vacancy postings

- In the ProQuest data set, the advertisements on a single page are all grouped in a single text field.
  - Our next task is to identify when one vacancy posting ends and a second posting begins. We use the following three-step rule to demarcate individual ads:
    1. **Zip codes and Addresses**
    2. **Ending phrases:** "send [...] resume", "submit [...] resume", "in confidence to", "affirmative employer","equal opportunity"
    3. **Job Titles**

Identifying the advertisement's job title

- Construct a list of one-word personal nouns, both in both singular and plural forms (e.g., engineer, clerk, manager) from the "Sample of Reported Job Titles" section in the O*NET website.
- Each line in the advertisement is defined as a job title if the following conditions are met:
  1. Words are all-capitalized or there are at most two words in a line.
  2. Contains at least one word from the list of one-word personal nouns.
- For example, with the word "clerk" and "clerks" in our list of one-word personal nouns, we can detect both "MUTUAL FUNDS CLERKS" and "PAYMENT CLERKS" without having to construct a full list of clerical occupations.

Deming and Kahn (2016) Skill Classification

cognitive	analytical, cognitive, critical thinking, math, problem solving, research, statistics
social	collaboration, communication, negotiation, presentation, social, teamwork
character	detail oriented, meeting deadlines, multi-tasking, time management
writing	writing
customer service	client, customer, customer service, patient, sales
project management	project management
people management	leadership, mentoring, people management, staff, supervisory
financial	accounting, budgeting, cost, finance, financial
computer	computer, software, spreadsheets

Trends in keyword frequencies (Spitz-Oener, 2006 class.)

	Frequency						Within
	60-64	65-69	75-79	80-84	90-94	95-99	Share
Nonroutine	4.06	4.47	4.93	4.77	5.36	6.42	0.68
Analytic	[3.95]	[4.22]	[4.34]	[4.54]	[4.93[	[5.54]	
Nonroutine	5.50	4.84	5.67	6.31	7.42	8.10	0.59
Interactive	[5.30]	[4.85]	[5.11]	[5.89]	[6.40]	[6.84]	
Nonroutine	1.75	1.71	1.92	2.03	2.20	2.15	0.89
Manual	[1.77]	[1.71]	[1.82]	[1.99]	[2.20]	[2.13]	
Routine	1.47	1.33	0.96	0.99	0.55	0.61	0.60
Cognitive	[1.51]	[1.38]	[1.15]	[1.20]	[1.00]	[1.00]	
Routine	0.45	0.38	0.28	0.18	0.05	0.03	0.92
Manual	[0.46]	[0.37]	[0.29]	[0.20]	[0.10]	[0.08]	

Trends in keyword frequencies (Deming Kahn, 2016 class.)

	Frequency						Within
	60-64	65-69	75-79	80-84	90-94	95-99	Share
Character	4.72	5.13	6.20	6.84	6.79	5.95	0.90
	[4.71]	[5.09]	[6.29]	[6.82]	[7.02]	[5.81]	(0.08)
Computer	1.05	1.29	1.90	2.23	3.80	4.50	1.09
	[1.05]	[1.30]	[1.89]	[2.15]	[4.00]	[4.82]	(0.03)
Customer	2.90	2.73	3.55	3.76	5.16	5.21	0.90
Service	[2.89]	[2.67]	[3.53]	[3.72]	[4.95]	[4.97]	(0.04)
Problem	1.31	1.29	1.19	1.30	1.81	2.01	0.73
Solving	[1.30]	[1.26]	[1.10]	[1.18]	[1.74]	[1.81]	(0.06)
Social	0.39	0.45	0.64	0.95	1.67	1.96	0.91
	[0.38]	[0.44]	[0.60]	[0.89]	[1.63]	[1.81]	(0.03)
Writing	0.47	0.42	0.47	0.52	0.87	0.80	0.90
	[0.46]	[0.40]	[0.39]	[0.46]	[0.77]	[0.76]	(0.05)

"Frequency" refers to the number of task-related keywords per thousand words.

Emulating O*NET's database

- Work styles (O*NET Elements 1C)
- Skills (O*NET Elements 2A and 2B)
- Knowledge requirements (O*NET Elements 2C)
- Work activities (O*NET Elements 4A)

Firpo, Fortin, and Lemieux (2011) construct their measures of information content, face-to-face contact, on-site job and decision-making from O*NET database (see their Appendix Table A.2).

Statistical Decomposition

- Definition (for a particular quantile)

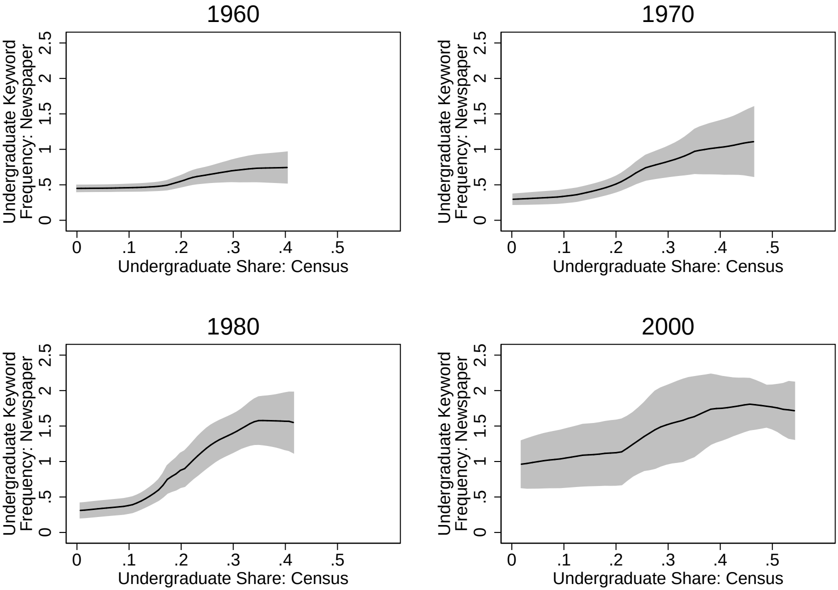
$$\mathrm{RIF}\left(w;\tau\right) \equiv w_{\tau} + \frac{\tau - \mathbf{1}\left(\mathrm{w} < \mathrm{w}_{\tau}\right)}{f\left(w_{\tau}\right)}$$

- Then the wage quantile can be written as τ

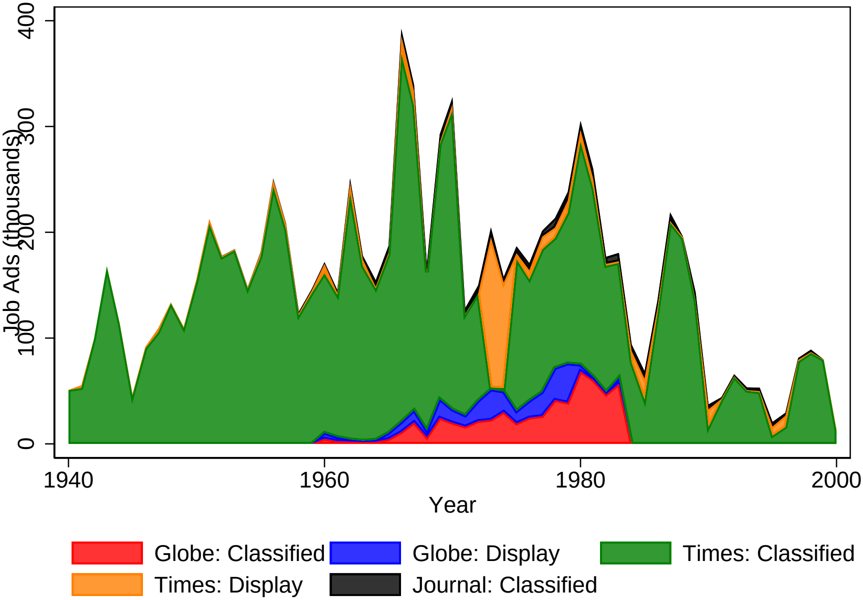
$$w_{\tau} = \int \mathbb{E}[\mathrm{RIF}\left(w;\tau\right)|X = x]dF_X(X)$$

- If we assume that $\mathbb{E}[\mathrm{RIF}\left(w;\tau\right)|X = x]$ is a linear function of X, we can use the Oaxaca-Blinder decomposition on $\mathrm{RIF}\left(w;\tau\right)$

Check on educational requirement



Number of Ads per Newspaper-Year



Words per Ad, by Newspaper-Year