［文献］Labussière, M., & Bol, T. (2024). Are occupations “bundles of skills”? Identifying latent skill profiles in the labor market using topic modeling. *OSF Preprints.* https://doi.org/10.31219/osf.io/5zwmt

麦山 亮太（学習院大学）

2024/07/17 Reading Circle for Inequality Studies

*掲載された論文ではないですが、とてもおもしろく、近いうちに competitive なジャーナルに掲載されると思うのでご容赦ください。

Abstract

Occupations are a central unit for understanding inequality in the labor market, yet we know little about why occupations matter. The existing literature often assumes that occupations are distinct bundles of skills, so skills are rarely conceptualized and measured independently of occupations. How does this limit our understanding of wage inequality? In this article, we use a unique dataset of millions of online job postings in the United Kingdom to capture the skill content of work at the job level and analyze its relation- ship to existing occupational classifications. While previous literature has often defined different skills as uni-dimensional and independent from each other, we propose a skill profile approach to capture the combination of skills that workers need on the job. Using topic modeling on highly detailed job skill requirements, we identify the skill profiles of job postings and analyze the extent to which they affect wages *within* or *across* occupational categories. Our results reveal substantial heterogeneity in skill content within occupations, and show that both job-level skills and occupations are key to explaining wage differentials across jobs. These findings challenge the often assumed role of occupations as bundles of skills, and offer new perspectives for analyzing labor market stratification.

背景および問い

社会学者と経済学者はともに職業を不平等を生み出す重要な要因として注目してきた。ここで職業は類似する業務（タスク）を行う仕事の集まりとして定義され、また職業は類似するスキルを持つ労働者の集合として理解されてきた。たとえば職業間の賃金格差が拡大したというとき、職業で求められるスキルに対するリターンが変化したことに原因が求められてきたし（Autor and Handel, 2013; Liu and Grusky, 2013）、職業移動が起こる要因も、職業はそれぞれ固有のスキルを必要とし、必要なスキルが類似していない職業への移動は起こりにくい（スキルが移動障壁を生み出す）と論じられてきた（Gathmann and Schönberg, 2010; Kalleberg and Mouw, 2018; Cheng and Park, 2020）。

いずれも職業が不平等を生み出す要因をそのスキルの水準に求めているものの、それは

理論的にそのように想定されているだけで、明示的に検証されていない。実際、職業がスキルのまとまりだ（=同一職業内のスキルは類似している）という暗黙の仮定は、経験的な証拠によって疑義が向けられている：

1. 異なる職業であっても、必要とされるスキルには重なりがある（Poletaev and Robinson, 2008; Gathmann and Schönberg, 2010; Cheng and Park, 2020）。
2. 同一職業でも仕事の内容はしばしば異なる（Yamaguchi, 2012; Autor and Handel, 2013; Cassidy, 2017; Freeman et al., 2020; Martin-Caughey, 2021）

もし職業が類似するスキルをまとめたものだという仮定が正しくないのだとすれば、われわれは労働市場における不平等をみるうえでのスキルや職業の役割を問い直さなければいけない。ゆえにこの問題は重要である。

そこで本研究は、職業は本当にスキルを束ねたもの（bundle of skills）なのか、もしそうでないなら、仕事（job）に必要となるスキルが職業とは無関係に賃金を説明する程度というのはどれくらいなのか、という問いに答える。

何が新しいのか

本研究は以下の 3 点で先行研究の問題を乗り越える。

1. 先行研究は限られた数のスキルを取り上げ、それらの個別の効果を明らかにするにとどまっていた。それに対して本研究は、つまりスキルを多次元的なものとして捉えたうえで、種類の異なるスキルの組み合わせ（skill profile）に注目する。
2. 先行研究は研究者・調査設計者があらかじめ決めたスキルを用いてきた。これに対して本研究は、データからボトムアップでスキルを分類する。
3. 先行研究は O*NET や DOT など、between-occupation の指標（同一職業であればスキルは同一となる）を利用してきた。これに対して本研究は between-occupation と within-occupation、双方の分散を活用して分析する。

方法

データ

分析には、Lightcast 社が収集している、イギリスのオンライン求人（job posting）情報を 51000 以上の仕事に関係するウェブサイトから収集したデータを用いる。ウェブ上に存在する求人をほぼすべて網羅できているといえる。職名（job title）、給与額、職業コード、32000 のキーワードからなるスキル要件が付されている。このデータは上記の目的に適している一方で、以下の限界もある。

- オンラインで公開されていない求人は捕捉できない（とくに小企業や零細企業など）
- 職業コーディングは機械学習によって行われており、誤分類があるかもしれない。

ただし誤分類しやすい職業は人間のコーダーにとっても分類が難しい職業である。

分析に使用するのは 2019 年分の、何らかのスキル要件を記載している求人情報データ（N≃6,300,000）から 10%無作為抽出（N=600,000）されたデータ。賃金の分析では、賃金（提示されている最低保障賃金）の情報があるケース（約 62%）を用いる。

分析 1：スキルカテゴリの抽出

分析手法：トピックモデル（biterm topic model）

- 各文書（各求人情報）は$t_1, \cdots, t_k$のトピックから生成されたものと考える。文書内のトピック出現確率p_{d_m,t_k}をすべて（Table 2 でいうと行方向）足すと 1 になる。
- 各トピックは出現確率の異なるさまざまな単語$w_1, \cdots, w_N$の集合として表される。トピック内の単語出現確率p_{t_k,w_N}をすべて（Table 3 でいうと行方向）足すと 1 になる。

文書のトピック分布とトピックの構造はいずれも同一文書内における単語の共起関係から推定される。ただし、文書が短い場合、推定が不安定になる。そこで、2 単語の組み合わせも bag-of-words に加える biterm topic model を用いて推定する。

	t_1	t_2	.	.	.	t_K
d_1	p_{d_1,t_1}	.	.	.	.	p_{d_1,t_K}
d_2	.	.				.
.	.		.			.
.	.			.		.
.	.				.	.
d_D	p_{d_D,t_1}	.	.	.	.	p_{d_D,t_K}

(a) Document-topic matrix, definition

	"Software Engineering"	"Management"	"Graphic Design"
Web developer	0.7	0.1	0.2
Software manager	0.5	0.4	0.1
Graphic designer	0.1	0.1	0.8

(b) Document-topic matrix, illustration

Table 2: Definition (a) and illustration (b) of the document-topic matrix.
Note: We note $d_1, ...d_D$ the D documents, $t_1, ...t_K$ the K topics, and p_{d_i,t_i} the probability that document d_i contains topic t_i. The distribution of each document over the topics adds up to one, i.e., $\sum_{k=1}^{K} p_{d_m,t_k} = 1$ for all topic d_m in $d_1, ...d_D$. The example is based on a fictional case with job postings for three job titles (Web developer, Software manager, Graphic designer) and three topics ("Software engineering", " Management", " Graphic Design").

	w_1	w_2	.	.	.	w_N
t_1	p_{t_1,w_1}	.	.	.	.	p_{t_1,w_W}
t_2	.	.				.
.	.		.			.
.	.			.		.
.	.				.	.
t_K	p_{t_K,w_1}	.	.	.	.	p_{t_K,w_W}

(a) Topic-term matrix, definition

	Budgeting	Creativity	Java	Adobe InDesign
"Management"	0.9	0.1	0.0	0.0
"Graphic Design"	0.0	0.4	0.0	0.6
"Software Engineering"	0.0	0.0	0.8	0.2

(b) Topic-term matrix, illustration

Table 3: Definition (a) and illustration (b) of the topic-term matrix.
Note: We note $t_1, ...t_K$ the K topics, $w_1, ...w_N$ the N words from the corpus vocabulary and p_{t_i,w_i} the probability that topic t_i contains word w_i. The distribution of each topic over the words adds up to one, i.e., $\sum_{n=1}^{N} p_{t_k,w_n} = 1$ for all topic t_k in $t_1, ...t_K$. The example is based on a fictional case with three topics ("Software engineering", " Management", "Graphic design") and a vocabulary of four skill words (Budgeting, Creativity, Java, AdobeInDesign).

結果

上記トピックモデルを用いて、Table 4 のとおり 19 個のトピック（以下、スキルカテゴリ）を抽出した。

	Latent skill category	**The 15 most relevant skills**, by decreasing order ($\lambda = 0.7$)
1	Project Management	budgeting, project management, planning, communication skills, stakeholder management, people management, building effective relationships, key performance indicators KPIs, quality management, staff management, problem solving, teamwork/collaboration, performance management, change management, organizational skills
2	Office administration and management	Microsoft Excel, administrative support, organizational skills, Microsoft Office, communication skills, detail-orientated, Microsoft Word, secretarial skills, typing, Microsoft Powerpoint, customer service, spreadsheets, data entry, Microsoft Outlook, general office duties
3	Communication and Interpersonal Abilities	communication skills, detail-orientated, teamwork/collaboration, customer service, organizational skills, problem solving, writing, English, verbal/oral communication, time management, Microsoft Excel, listening, French, German, Spanish
4	Sales and Business Development	sales, business development, sales management, sales goals, account management, customer service, product sales, communication skills, business-to-business, prospective clients, building effective relationships, customer contact, client-base retention, telesales, teamwork/collaboration
5	Caregiving and Support Services	teaching, working with patient and/or condition: mental health, care planning, childcare, dementia knowledge, communication skills, nursing home, autism diagnosis treatment care, creativity, planning, English, home management, social services, learning disability, organizational skills
6	Customer Service and Retail Operations	cleaning, customer service, communication skills, cooking, retail industry knowledge, food safety, teamwork/collaboration, organizational skills, stock control, cash handling, food preparation, store management, English, housekeeping, detail-orientated
7	Digital Marketing and Content Strategy	social media, marketing, digital marketing, creativity, marketing management, Google Analytics, market strategy, content management, copy writing, editing, writing, email marketing, budgeting, social media tools, research
8	Financial Operations	accounting, finance, account reconciliation, balance sheet, Microsoft Excel, budgeting, VAT returns, financial reporting, financial accounting, bank reconciliation, accruals, statutory accounts, communication skills, detail-orientated, variance analysis
9	Web Development and Software Engineering	Javascript, Microsoft hash, software development, Java, NET, SQL, Git, software engineering, DevOps, Scrum, Python, active server pages ASP, ASP.NET, continuous integration CI, AngularJS
10	Logistics and Supply Chain Management	procurement, purchasing, supply chain knowledge, logistics, supply chain management, enterprise resource planning ERP, SAP, procurement contracts, planning, manufacturing resource planning MRP, contract management, Microsoft Excel, key performance indicators KPIs, material requirement planning MRP, communication skills

Continued on next page

Table 4: Description of the 19 latent skill categories obtained from the biterm topic model, using the 15 most relevant skills by decreasing order.

	Latent skill category	**The 15 most relevant skills**, by decreasing order ($\lambda = 0.7$)
11	Facility Maintenance	plumbing, preventive maintenance, predictive preventative maintenance, electrical work, carpentry, painting, HVAC, boilers, wiring, communication skills, hand tools, heating systems, customer service, cleaning, emergency lighting
12	Healthcare and Patient Care	patient care, surgery, anaesthesiology, communication skills, working with patient and/or condition: trauma, dentistry, orthopaedics, xrays, critical care, paediatrics, primary care, rehabilitation, gynecology, blood pressure measurement, urology
13	Business strategy	risk management, communication skills, business development, research, due diligence, project management, building effective relationships, analytical skills, economics, accounting, Microsoft Excel, asset management industry knowledge, insurance underwriting, teamwork/collaboration, planning
14	Engineering and Technical Expertise	AutoCAD, mechanical engineering, engineering design and installation, commissioning, engineering design, calculation, civil engineering, project management, mechanical design, Revit, systems engineering, simulation, communication skills, electronics industry knowledge, planning
15	Manufacturing and Engineering	computer numerical control CNC, machining, engineering drawings, welding, lathes, manufacturing processes, quality assurance and control, ISO9001standards, MIG and TIG welding, lean manufacturing, milling cutters, machine tools, computerised numerical control lathes, quality management, problem solving
16	Data Management and Analysis	SQL, python, machine learning, data science, tableau, data warehousing, Microsoft Power BI, extraction transformation and loading ETL, data analysis, business intelligence, big data, Apache Hadoop, SQL server reporting services SSRS, data architecture, Microsoft SQL server integration services SSIS
17	Technical Support and Troubleshooting	Microsoft Active Directory, VMware, Windows Server, Cisco, troubleshooting, ITIL, domain name system DNS, Microsoft Windows, IT support, wide area network WAN, technical support, dynamic host configuration protocol DHCP, transmission control protocol/internet protocol TCP/IP, Linux, Microsoft Exchange
18	Graphic Design and Creative Media	Adobe Photoshop, Adobe InDesign, Adobe Acrobat, Adobe Creative Suite, Adobe Illustrator, creativity, graphic design, Adobe After Effects, digital design, typesetting, animation, editing, creative design, detail-orientated, teamwork/collaboration
19	Scientific Research and Laboratory Work	research, chemistry, biology, clinical trials, biotechnology, biochemistry, molecular biology, experiments, clinical research, cancer knowledge, oncology, drug development, bioinformatics, high performance liquid chromatography HPLC, drug discovery

Table 4 Continued: Description of the 19 latent skill categories obtained from the biterm topic model, using the 15 most relevant skills by decreasing order.
Note: We use the *relevance* of the skills as defined by Sievert and Shirley (2014) to account for the fact that some skill requirements (e.g. communication skills) are listed much more frequently than others, which mechanically increases their probability of belonging to any latent skill category. The parameter λ determines the weight given to the probability of a term under a given topic relative to its marginal probability in the corpus (Sievert and Shirley, 2014, 66). We observed that $\lambda = 0.7$ provides optimal interpretability, but note that the latent topics are easily interpretable without this correction (see Table S3 in the Supplementary Materials).

スキルカテゴリのなかには、一般的なスキルからなるもの（1 Project Management; 3 Communication and Interpersonal Abilities）もあれば、仕事特有のスキル（10 Logistics and Supply Chain Knowledge）、特定の技術（15 Manufacturing and Engineering）、ノウハウ（12 Healthcare and Patient Care）、特定のソフトウェアや ICT に関するスキル（14 Engineering and Technical Expertise）を含むものもあり、多様である。

Figure S15 は、文書全体における各トピックの分布を見たものである。各求人（job）が単一のトピックのみを含む（もしそうならば 0 と 1 に分布が集中する）のかというと、そうではない。ほとんどの文書は複数のトピックを含んでおり、job によって多様なスキルカテゴリ（skill profile）が要求されている。

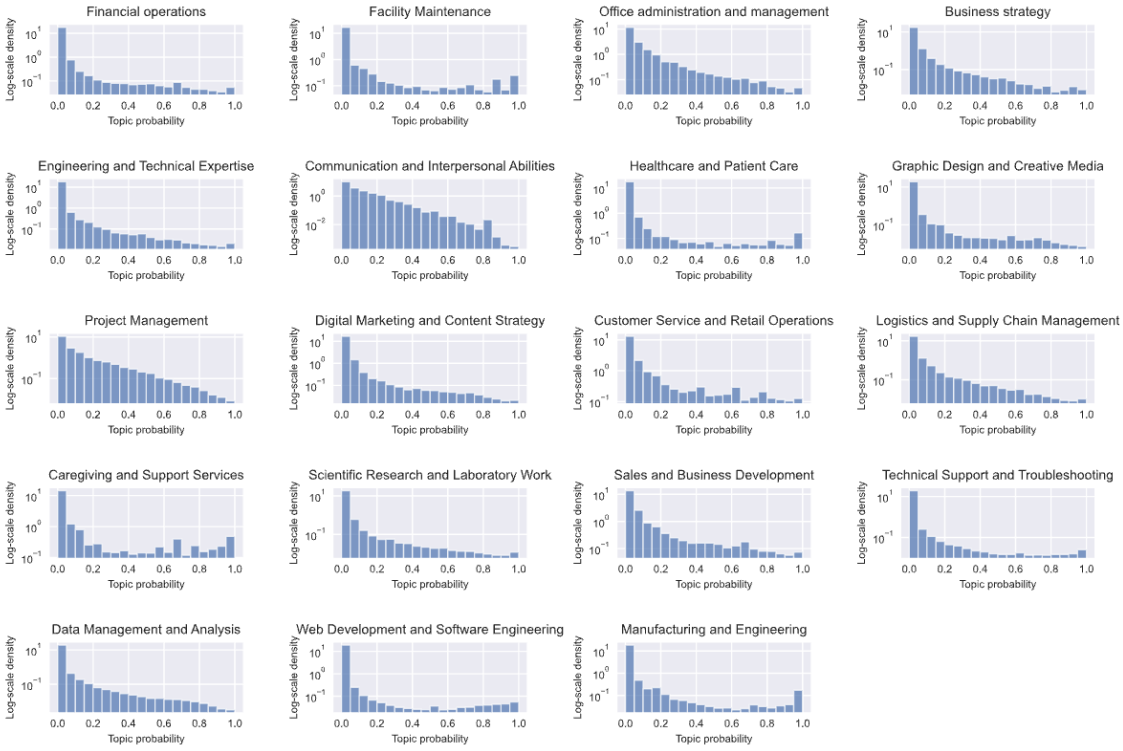


Figure S15: Probability distribution of each latent skill category within the sampled job postings, with the standard log-scale density function. *Note: Y-axes differ between latent skill categories.*

分析 2：職業間の skill profile の類似性

分析手法：Maximum mean discrepancy (MMD) distance

カーネル法にもとづき 2 つの分布の重なりの程度を比較する方法。直感的には、2 つの確率分布が完全に重なっていれば 0、まったく重なっていなければ 1 を取る（Duncan's dissimilarity index みたいな？）。

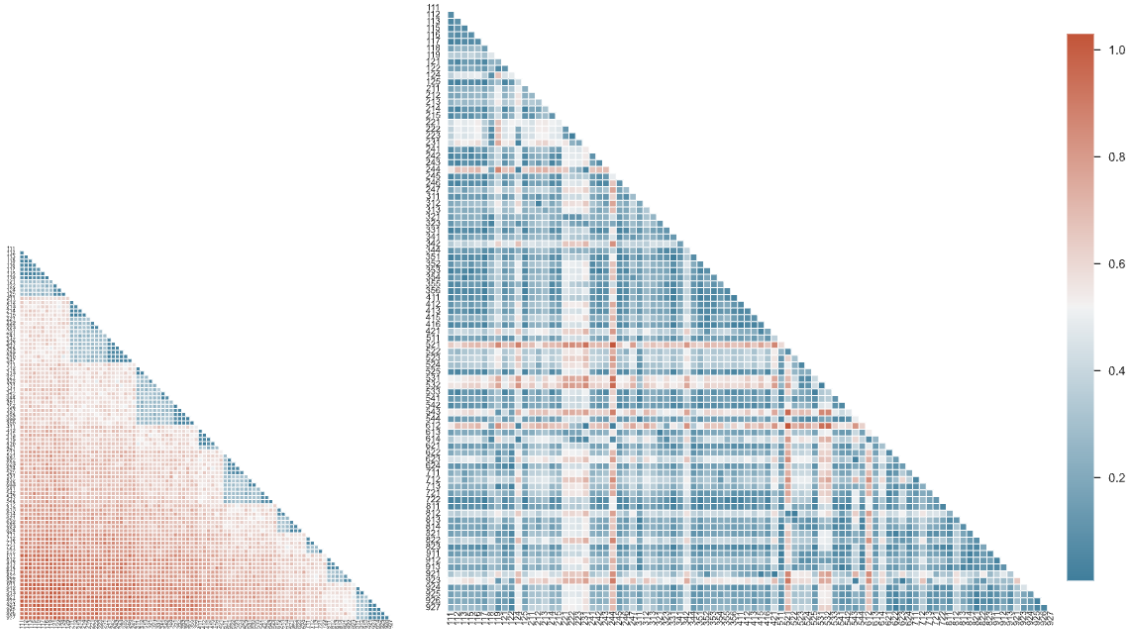
各 job のトピック分布を職業ごとに平均すると、職業のトピック分布が求まる。任意の 2 つの職業間でトピック分布（skill profile）を比較して、MMD を求める。MMD が職業間で大きく異なっていれば、職業内の類似性は職業間の類似性よりも高いとみなせる。

結果

3-digit SOC group で MMD を求めたのが Figure 6 (b)。もし、大分類内で skill profile が似て

いて、大分類間で異なっているのであれば、結果は(a)のようになるはずである。しかし実際には(b)のように、skill profile が大分類内で似ているとはいえない。また、多くの職業は skill profile はかなり似通っている（MMD が低い）。4-digit で分けても結果は同様である。したがって、EGP や micro class が理論的に想定してきたような同一職業内のスキルの類似性があるとはいえない。

また、異なる職業（大分類）間でスキルの類似性が高い職業も存在する。たとえば、清掃員（923）と家政婦（623）等が挙げられる。



(a) Simulated heat map　　(b) Empirical heat map

Figure 6: Simulated and empirical heat maps for the MMD distance between minor groups.
Note: The x- and y-axes are labeled from the coding in the SOC occupational classification. The simulated heat map (a) represents the theoretical expectation that minor occupations in the same sub-major group should be closer to each other than to minor occupations in other minor groups. It is based on a simulation with normal Gaussian distributions and is for illustrative purposes only. In the empirical heat map (b), each cell indicates the maximum mean discrepancy pairwise distance between minor groups based on their representation in the latent skill space. We use a kernel bandwidth of $\lambda = 1.49$.

分析 3：職業と skill profile の賃金に対する予測力

分析手法：回帰分析

求人に記載された時給の対数値を従属変数とする回帰分析で、決定係数を比較。

結果

- Skill profile（個々の求人のトピック確率 ＝ 記載されたスキル）は賃金をかなりの程度予測する。それだけでなく、職業を統制してもなおかなりの程度予測力を高める。
- 職業とスキルの賃金に対する予測力は代替的というよりは補完的である。つまり、

両者は相関するものの、それぞれが独自の効果を持っている。

		Model (1) *Skill profiles*		**Models (2)** *SOC coding*		**Models (3)** *(1) and (2) combined*	
		R^2	Df	R^2	Df	R^2	Df
		0.301	18				
SOC level	Major groups			0.289	8	0.403	26
	Sub-major groups			0.320	24	0.414	42
	Minor groups			0.356	89	0.432	107
	Unit groups			0.396	368	0.460	386

Table 7: Adjusted coefficient of determination R-squared (R^2) and degrees of freedom (Df) for the prediction of the logged minimum hourly wage included in job postings, comparing the models (1)-(3) defined on p.28.
Notes: Degrees of freedom represent the number of effective parameters in the models.

限界

- 本データは必要なスキルを提示している一方で、スキルや能力のレベルについては記載していない。
- 雇用主が提示した仕事に必要なスキルの内容が、実際に労働者が行っていることと一致するとは限らない。たとえば、持っていて当たり前と思われているスキルはわざわざ求人に書かれていないかもしれない。
- 提示されている賃金が実際に労働者が受け取る賃金と一致するとは限らない。したがって労働者間に実際存在する賃金の不平等とは一致しないかもしれない。
- 今回のデータドリブンに求めたスキルの分類はあくまで今回のデータから求められたもので、外的妥当性がどの程度あるかは今後の課題である。

議論

職業は類似するスキルをまとめたもの（bundle of skills）ではない。多くの仕事は異なる種類のスキルを必要とし、またそれは一般的なスキルから特殊な知識が混じったものである。加えて、職業内というよりは職業間で似通ったスキルを要するものも多い。

職業とスキルは互いに異なる概念である。それは職業とスキルがそれぞれ独立に賃金と関係していることからも明らかである。

- 職業とスキルの研究への含意：職業はスキルの代理指標にはならず、何か異なるものを捉えている可能性がある。
- （世代内）社会移動研究への含意：職業間でのスキルの類似性の高さは、従来考えられているよりも職業間移動はしやすい可能性を示唆している。
- 職業と不平等の研究への含意：Skill profile を統制しても、職業自体はなお賃金を予測する。したがって職業が不平等を生み出すメカニズムはスキルだけではない社会

的要因（社会的閉鎖など）によるものと考えられる。

- 制度的要因の重要性：今回の研究はイギリスという教育と職業の結びつきが弱い社会で得られた結果であり、このような文脈では within-occupation のスキルのばらつきが大きくなっている可能性がある。他の社会での研究がさらに必要である。

コメント

- リサーチクエスチョンが明確で、理論的に自明とされている前提を問い直すというフレームなのがとても良い。分析や結論は手堅く方法も適切で、分析についてとくに批判すべき点は見当たらなかった。
- トピックモデルを使って個体のトピック確率を抽出して変数を作るというのは麦山・西澤（2016）のほか、社会調査の自由記述データ（横山智哉，2019,「トピックモデルを用いた政治的会話の構造の推定」『理論と方法』34(2): 206–219）などでも見たことがあるので、適切なデータがあれば使いどころがあるかもしれない。方法の説明や応用例の紹介として Grimmer, Jastin, Margaret E. Roberts, and Brandon M. Steward. 2022. *Text as data: A new framework for machine learning and the social sciences.* Princeton University Press.など。
- 序論が長く（13 paragraphs)、ここで論文の意義や先行研究との違い、データ、分析結果についてすべて説明し、Background で背景について補足するというスタイル。経済学や、社会学でも AJS などではこういうスタイルの論文は多いが、今後こういうスタイルが流行っていくのだろうかと思った。
- 採用時点で持っていることが期待されるスキルと、採用後に身につけることを期待されているスキルの間に隔たりがあるとすれば、この論文の主張は必ずしも正しくない点が限界として挙げられる。たとえば、日本の新卒採用のように採用時点では一般的な employability を持っていることが期待され、実際に仕事に使うスキルは企業内の OJT で身につけていくような場合であれば、採用時点で期待されるスキルと労働者が持っているスキルは一致しないと思われる。
- 求人時点ではあくまで（応募者数を確保するために）ざっくりとしたスキルだけを記載し、実際にはもっと細かな要件を要求することがあるかもしれない。もしそうだとしたら、実際にはもっと職業ごとに固有のスキルがあるということになる。
- ここで測定されているのはスキルだが、職業の定義は類似するタスクのまとまりであるため、両者がとくに断りなく入れ替わっているのが気にならないでもなかった。ただし、神林（2016: 301）は「使用者が設計するタスクと、被用者の持つスキルはコインの裏表のような関係にあり、中長期的には一致していると考えられる」と述べている。