A Task-based Approach to Constructing Occupational Categories with Implications for Empirical Research in Labor Economics

by

Julia Manzella
U.S. Census Bureau

Evan Totty
U.S. Census Bureau

Gary Benedetto
U.S. Census Bureau

CES 19-27 September, 2019

Abstract

Most applied research in labor economics that examines returns to worker skills or differences in earnings across subgroups of workers typically accounts for the role of occupations by controlling for occupational categories. Researchers often aggregate detailed occupations into categories based on the Standard Occupation Classification (SOC) coding scheme, which is based largely on narratives or qualitative measures of workers' tasks. Alternatively, we propose two quantitative task-based approaches to constructing occupational categories by using factor analysis with O*NET job descriptors that provide a rich set of continuous measures of job tasks across all occupations. We find that our task-based approach outperforms the SOC-based approach in terms of lower occupation distance measures. We show that our task-based approach provides an intuitive, nuanced interpretation for grouping occupations and permits quantitative assessments of similarities in task compositions across occupations. We also replicate a recent analysis and find that our task-based occupational categories explain more of the gender wage gap than the SOC-based approaches explain. Our study enhances the Federal Statistical System's understanding of the SOC codes, investigates ways to use third-party data to construct useful research variables that can potentially be added to Census Bureau data products to improve their quality and versatility, and sheds light on how the use of alternative occupational categories in economics research may lead to different empirical results and deeper understanding in the analysis of labor market outcomes.

---

[*] Disclaimer: Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. All results have been reviewed to ensure that no confidential information is disclosed. Data in this report based on confidential data were reviewed and released by the Census Bureau Disclosure Review Board, approval CBDRB-FY19-CED001-B0024 and CBDRB-FY19-501.

Contact: Manzella - Corresponding author. Affiliation: Center for Economic Studies, U.S. Census Bureau. Address: U.S. Census Bureau; 4600 Silver Hill Road; Washington, DC 20233. Manzella: julia.ann.manzella@census.gov. Totty and Benedetto - Affiliation: Center for Enterprise Dissemination, U.S. Census Bureau. Address: U.S. Census Bureau; 4600 Silver Hill Road; Washington, DC 20233. Benedetto: gary.linus.benedetto@census.gov; Totty: evan.scott.totty@census.gov

Acknowledgements: The authors appreciate helpful comments on earlier drafts of this paper from peer reviewers, presentations at 2018 BLS-Census Annual Workshop, Census Economic Research Brown Bag, 2018 Southern Economic Association Conference, and 2019 Society of Labor Economics (SOLE) Conference.

A Task-based Approach to Constructing Occupational Categories with Implications for Empirical Research in Labor Economics

1 Introduction

The aim of this paper is to provide a quantitative approach to aggregating detailed occupations that utilizes data on tasks performed and is grounded in theory. We examine whether, and to what extent, using alternative occupational categories in economics research may lead to different empirical results and deeper understanding in the analysis of labor market outcomes.

There exists a large literature that examines differences in earnings across subgroups of workers wherein one typically accounts for the role of occupations by controlling for occupational categories (e.g., O'Neill, 1990; Hirsch, 2004; Blau & Kahn, 2017). In a recent survey article in the *Journal of Economics Literature*, Blau and Kahn (2017) provide such an example in their analysis of the gender wage gap. Researchers typically aggregate detailed occupations into categories based on the Standard Occupation Classification (SOC) coding scheme. At the two-digit level, the SOC codes aggregate detailed occupations into Major Groups under the assumption that workers within these occupational categories perform similar work tasks.

There is also a body of literature that examines returns to worker skills. In estimating wage differentials for caring work, Hirsch and Manzella (2015) merge O*NET job descriptors with worker-level data from the Current Population Survey (CPS) to construct comprehensive, continuous measures of "caring" across all occupations. Several interesting findings were uncovered when they analyzed the merged data. For example, while teachers and homebuilders belong to two different major occupation groups derived from two-digit SOC codes, both rank highly in the amount of "developing/teaching others", which is one type of "caring."[1]

We propose two quantitative, task-based approaches that can potentially provide a different way of grouping numerous detailed occupations into several, broader occupation groups that may be more economically meaningful than SOC Major Groups (i.e., two-digit SOC codes). More homogenous skill/task groupings would be very relevant for empirical research and therefore may be valuable additions to Census Bureau data products such as the Survey of Income and Program Participation (SIPP) Gold Standard File (GSF), which is made available to external researchers by creating a non-disclosive, synthetic version of the data called SIPP Synthetic Beta (SSB).[2] Even if the occupational categories formed from quantitative analyses of O*NET job descriptors are well-aligned with the occupational categories formed from using two-digit SOC codes, undertaking such an investigation would be useful in offering validation of the existing SOC-based approach in empirical research.

Our study is the first to employ factor analysis with a large, rich set of O*NET job descriptors to create a few latent skill/task factors used to construct occupational categories,

---

[1] Teachers are grouped into the two-digit 2000 SOC code 25-0000 "Education, Training, and Library Occupations" while homebuilders belong to the two-digit 2000 SOC code 49-0000 "Installation, Maintenance, and Repair Occupations". Detailed 2000 SOC codes for various teacher occupations are 25-1000 through 25-9099, while the detailed 2000 SOC code for "Manufactured Building and Mobile Home Installers" is 49-9095. There are no differences between the 2000 SOC and 2010 SOC codes for these detailed occupations or major groups.

[2] U.S. Census Bureau Gold Standard File consists of data from respondents on the SIPP for panels 1984-2008 linked with tax and benefit data from the Internal Revenue Service (IRS) and Social Security Administration (SSA). It is not feasible to synthesize detailed occupations due to the complexity of modeling relationships among many variables. Therefore, the Census Bureau summarizes detailed occupation information using occupational categories and synthesizes the categories. Outside researchers can have their SSB-based results validated on non-synthetic data. More information is available here: https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html.

with the goal of achieving greater homogeneity of skills and tasks within each category.[3] More specifically, we estimate five latent skill/task factors and their loadings, then we predict factor scores for each latent factor in each detailed occupation. Next, detailed occupations are classified into (potentially) thirty-two occupational categories depending on whether a detailed occupation has a high or low predicted factor score value in each of the five latent skill factors (i.e., $2^5 = 32$ categories), using the median factor score value as the high-low threshold for a skill factor.

2 Background on the Standard Occupation Classification (SOC) system

Beginning with the implementation of the 2000 SOC, all Federal statistical agencies producing occupational data sources need to include SOC codes that classify workers and jobs into standardized detailed occupations and aggregate occupation groups.[4] For the U.S. Census Bureau, Census occupation codes are largely determined by collectability from household surveys. While the Census Occupation Code list is based on the SOC, the mapping between Census occupation codes and SOC codes may be one-to-one or one-to-many.[5]

Roughly every ten years, there is a formal SOC revision process where public comment is collected through Federal Register notices. It is possible that special interest groups may request changes based on their own needs (e.g., tax-reporting or compensation determined through collective bargaining) rather than based on a broader interpretation of the occupation.

[3] Dey and Lowenstein (2019) have recently done related and interesting work. One important difference between their work and ours is that their goal is to use O*NET tasks to explain wages, resulting in an occupation aggregation scheme. The aim of our study is to devise an aggregation method that is solely based on tasks and is independent of wages.

[4] More information on the historical background of the 2000 SOC can be found in a current version of the SOC manual at https://www.bls.gov/soc/2018/soc_2018_manual.pdf.

[5] Census Bureau Occupation Code Lists provide a mapping between Census occupation codes and SOC codes, and they are available at https://www.census.gov/topics/employment/industry-occupation/guidance/code-lists.html.

Next, interagency workgroups review public comments and provide recommendations to the SOC Policy Committee who makes final determinations. Public input is subject to careful review and consideration in accordance with the SOC classification principles and coding guidelines.[6]

We provide a recent example to illustrate how public comment could potentially influence SOC-based occupational categories and to highlight the importance of using quantitative analyses of task measures to validate SOC groupings. In regards to a Federal Register notice for the 2018 SOC revision process, respondents discussed why the SOC code for the detailed occupation, Police, fire and ambulance dispatchers (43-5031), should be included in the 2-digit SOC occupational category, Protective services occupations (33-0000), instead of in the 2-digit SOC occupational category, Office and administrative support (43-0000). Their fundamental argument was that public safety dispatchers typically spend more time performing the types of tasks and responsibilities utilized by workers in other occupations within the Protective service category (e.g., first-responders) than they do performing the types of tasks and responsibilities utilized by workers in other occupations within the Office and administrative support category (e.g., telephone operators and various clerks).

3 Data

3.1 Job Descriptor Data

The Occupational Information Network (O*NET) was developed under the sponsorship of the U.S. Department of Labor/Employment and Training Administration (USDOL/ETA), and it is the nation's primary source of occupational information. The O*NET database contains a rich

[6] The most recent 2018 SOC classification principles and coding guidelines are available in the 2018 SOC User Guide at https://www.bls.gov/soc/2018/soc_2018_class_prin_cod_guide.pdf. Documentation for 2000, 2010, and 2018 SOC vintages is available at https://www.bls.gov/soc/.

set of job descriptors. We use 144 detailed job descriptors from the O*NET database version 15.0.[7] The O*NET job descriptors provide quantitative measures of worker characteristics (abilities and workstyles), worker requirements (basic and cross-functional skills), and occupational requirements (work context and generalized work activities) at the detailed occupation level. Each descriptor is associated with a scale that provides a quantitative measure based on ratings from occupational experts and job incumbents (i.e., workers) surveyed throughout the U.S. The ratings scales (Level, Importance, Context) indicate the degree to which a particular descriptor is needed in the occupation. We rescaled all O*NET variables on [0, 1].[8]

O*NET descriptor values are assigned SOC codes at the detailed occupation level. Our employment data will come from the GSF and it uses 2002 Census occupation codes, which are based on the 2000 SOC, but due to collectability issues sometimes collapses detailed occupations into broad occupations. As a result, we have an O*NET data set with 485 detailed/broad Census occupation codes. In the instances when more than one 2000 SOC code was paired with one 2002 Census occupation code, we assigned the mean O*NET descriptor value to the detailed/broad Census occupation code. For example, the 2002 Census Occupation Code 2300 is equivalent to the SOC broad occupation, Preschool and Kindergarten Teachers (SOC 25-2010), which collapses two detailed occupations Preschool Teachers, except Special Education (SOC 25-2011) and Kindergarten Teachers, Except Special Education (SOC 25-2012).

[7] See Appendix B: Data Appendix for a detailed discussion of how our O*NET dataset was created.

[8] Possible original range of values for the different ratings scales are as follows: Level on [0, 7]; Importance on [1, 5]; and Context on [1, 5]. The rescaling formula uses the original rating value, and the lowest and highest possible rating values where the rescaled value = (original-lowest) / (highest-lowest).

In our O*NET dataset, the O*NET descriptor values for the two detailed occupations (SOC 25-2011 and 25-2012) are averaged to obtain the mean O*NET descriptors values for Preschool and Kindergarten Teachers (SOC 25-2010).

4 Methodology

We develop two task-based approaches for constructing occupational categories that provide alternatives to SOC Major Groups (i.e., two-digit SOC codes). We refer to these two, more general, alternatives to the SOC-based approach as our structured and unstructured task-based approaches. We discuss each task-based approach in detail in sections 4.1 and 4.2 below.

We employ factor analysis in both of our task-approaches. We use factor analysis because it allows us to extract a small number of latent skill/task factors that explain a large amount of the variation in the observed O*NET descriptor rating variables.[9] We also gave consideration to the comprehensibility, versatility and relevance of factor analysis since we were looking for an approach that would appeal to both scholars and general users of Census Bureau data products. Factor analysis is commonly used by researchers to reduce dimensionality in skills and tasks, and the resulting factors are interpretable. The interpretability of the latent factors makes it easier to interpret the resulting task-based occupation groupings and to evaluate differences between task-based and SOC-based groupings. Furthermore, we are interested in developing a way of modeling occupation that could be useful for imputation or synthesis, which is facilitated by having five variables (i.e., latent skill/task factors) each with continuous values.

[9] An alternative approach to using factor analysis to reduce dimensionality is, for example, optimizing an objective function subject to constraints, and such an approach would have to have important differences from the implicit optimization found in the factor analysis method.

We provide a brief overview of the exploratory factor analysis approach we use in our task-based approaches to constructing occupational categories. A comprehensive summary of exploratory factor analysis can be found in Fabrigar et al. (1999). The first step in performing factor analysis involves estimating the factor loadings. In our context, factor loadings represent the O*NET descriptors on which the latent factors load most strongly. These can be interpreted as the regression coefficients that would be obtained by regressing the latent factors on the O*NET descriptors. The next step involves estimating the factor scores. A factor score for a detailed occupation is based on the factor loading for each O*NET descriptor and the ratings value of each O*NET descriptor. Predicted factor scores provide estimates of the extent to which the given occupation requires each latent skill factor. Note that we use the Bartlett approach to generate the factor scores, which is based on the product of the factor loading matrix, the inverse of the data covariance matrix, the observed descriptor values, and a correction term for bias in the factor means (Bartlett, 1937). Finally, determining the number of factors to be extracted is typically done by examining multiple criteria although no approach is considered to be perfect. We choose five factors for reasons discussed in section 4.2.

4.1 Structured Task-based Approach

Essentially, our first task-based approach to creating occupational categories provides a structure, based on seminal work by Autor, Levy, and Murnane (2003), for which the O*NET descriptors are loaded to form each latent skill/task factor. We arrange the O*NET descriptors into five task groups. Then we load only the subset of descriptors from a particular task group and extract only the first factor to estimate the corresponding latent skill/task factor. This process is repeated separately for each of the five groups of O*NET decsriptors. In contrast, our unstructured approach loads the entire set of O*NET descriptors (109 or 144) and extracts the first five factors to estimate the latent skill/task factors.

Autor, Levy, and Murnane (2003) examine how computerization changes job skill demands. They assume computer capital and labor are perfect substitutes in performing routine tasks and worker self-selection among occupations clears the labor market. In their empirical analysis, Autor, Levy, and Murnane (2003) select a subset of task variables from the U.S. Department of Labor's Dictionary of Occupational Titles (DOT)—the precursor to O*NET—to measure non-routine cognitive tasks, routine cognitive tasks, routine manual tasks, and non-routine manual tasks.[10]

In our structured approach, we select a subset of O*NET descriptor variables to measure non-routine interactive tasks (NRI), non-routine analytical tasks (NRA), routine cognitive tasks (RC), routine manual tasks (RM), and non-routine manual tasks (NRM). We employ two methods for selecting O*NET descriptor variables. Method 1 allows a descriptor to be placed into one and only one task group, yielding a set of 109 O*NET variables.[11] We refer to this structured factor analysis approach as Method 1-109. Method 2 allows a descriptor to be placed into multiple task groups, yielding a set of 144 O*NET variables.[12,13] This set of 144

[10] Non-routine cognitive tasks are measured by two variables: one for interactive skills and one for analytical skills; routine cognitive tasks are measured by one variable for adaptability to work requiring set limits, tolerances, or standards; routine manual tasks are measured by one variable for finger dexterity; and non-routine manual tasks are measured by one variable for eye-hand-foot coordination.

[11] In Method 1, non-routine interactive tasks are measured by 48 variables, non-routine analytical tasks are measured by 26 variables, routine cognitive tasks are measured by 12 variables, routine manual tasks are measured by 12 variables, and non-routine manual tasks are measured by 11 variables.

[12] In Method 2, non-routine interactive tasks are measured by 73 variables, non-routine analytical tasks are measured by 47 variables, routine cognitive tasks are measured by 20 variables, routine manual tasks are measured by 29 variables, and non-routine manual tasks are measured by 51 variables.

[13] We also apply Method 2 to the subset of 109 O*NET variables in order to check the sensitivity of our analysis, results are not shown due to space limitations. The Method 2-109 approach produces 22 non-empty occupational categories.

O*NET descriptor variables are comprised of 96 variables belonging to only one group that are also used in Method 1-109; 13 variables belonging to multiple groups that are also used in Method 1-109; and 35 variables belonging to multiple groups that were not used in Method 1-109. We refer to this structured factor analysis approach as Method 2-144.

In these structured approaches, we load only the subset of the full set of (109 or 144) O*NET descriptor variables corresponding to the latent skill/task group (i.e., NRI, NRA, RC, RM, or NRM) for a particular method and perform factor analysis. We extract only the first factor without rotation and its estimated loadings, and then we predict factor scores. We repeat this estimation procedure separately for each of the five latent skill/task groups. Next detailed occupations are classified into occupational categories depending on whether a detailed occupation has a high or low predicted factor score value in each of the five latent skill/task factors (2^5 = 32 potential occupational categories), using the median factor score value as the high-low threshold for a latent factor.[14]

4.2 Unstructured Task-based Approach

In our second task-based approach, we do not assume any direct relationship between the descriptors and the factors, nor do we place any direct interpretation on the factors extracted via factor analysis based on their construction. In this unstructured approach, we first perform factor analysis on the entire set of O*NET descriptors (109 or 144). We then extract the first five factors. Five factors were chosen for a few reasons. First, we want to have a number of

[14] We also employed an optimization routine that minimized the mean square error (where the ideal is 16 detailed occupations per occupational category) to determine the high-low threshold for each of the five latent skill factors. This did not improve the distribution of detailed occupations among the groups, a concern which we discuss in the Results section of the paper.

occupational categories that is comparable to the number of two-digit SOC categories (i.e., targeting a total of twenty to thirty categories). Second, we want be consistent with the theoretical underpinnings that are widely used by researchers (i.e., factors in the spirit of Autor, Levy, and Murnane (2003)). Moreover, the first five factors account for 78.6% of the variance in tasks and by the fifth factor the marginal percentage of variance accounted for is down to 2.5%.[15]

Next, we perform an oblique oblimin rotation of the factors, which allows the factors to be correlated. This approach seems appropriate given the fact that examples of potential latent skill/task factors could be expected to be correlated. For example, one might expect that occupations requiring more analytical work would often require less manual labor. The rotation of the factors can also ease interpretation of the factors when analyzing the factor loadings (Fabrigar et al., 1999). In the results discussed below, the factor loadings for each factor are used to interpret the factors.

Then, we predict factor scores for each detailed occupation. Lastly, detailed occupations are classified into occupational categories depending on whether a detailed occupation has a high or low predicted factor score value in each of the five latent skill/task factors, using the median factor score value as the high-low threshold for a latent skill/task factor.

4.3 Comparing Task-based Approaches to SOC-based Approach

We examine similarities and differences between task-based occupational categories and SOC-based occupational categories. Since our primary goal is to create more homogenous skill/task

[15] Appendix Table A1 shows the eigenvalues for the top 20 latent factors derived from the unstructured factor analysis using 144 O*NET descriptors.

groups, we start by comparing occupation distance measures. For every detailed occupation, we calculate the distance in tasks between the detailed occupation and its group mean. The intuition behind studying the occupation distance of a detailed occupation from its group mean is that if an approach does a good job of grouping detailed occupations that have similar tasks, then this distance measure should be lower than the distance measure of an alternative approach. The literature has typically used an Euclidean occupation distance measure (e.g., Robinson 2018). Equation 1 provides the primary Euclidean occupation distance measure in tasks that we employ.[16]

$$edistA = \sqrt{\left(t_{1,g,j} - \bar{t}_{1,g} \right)^2 + \left(t_{2,g,j} - \bar{t}_{2,g} \right)^2 + \cdots + \left(t_{I,g,j} - \bar{t}_{I,g} \right)^2} \qquad (1)$$

where $t_{i,g,j}$ is task *i* for detailed occupation *j* in occupation group *g,* and $\bar{t}_{i,g}$ is the mean value of task *i* for occupation group *g* (i.e., the group mean). Tasks are indexed by I = 1, 2, …, I and I is either 109 or 144 depending on the approach. Occupational categories or groups are indexed by g = 1, 2, …, G and G varies by approach. Detailed occupations are indexed by j = 1, 2, …, J_g and J_g varies by occupation group.

Additionally, we average the occupation distance across the number of detailed occupations within an occupational category to obtain an average within-occupation group distance measure. We also average the occupation distance across all detailed occupations to obtain a grand average occupation distance measure. The grand average occupation distance is used to summarize the occupation distances for an approach so that two approaches can be compared by a single quantitative measure. The approach with the lowest grand average occupation distance has the greatest homogeneity of tasks, and, thus, it is the most compelling way to construct occupational categories among the alternative approaches studied.

[16] In the subsequent analyses section, we discuss using a Mahalanobis occupation distance measure instead.

5 Results

5.1 Interpretation of factors from the unstructured approach

Tables 1 through 5 show the descriptors on which the five factors extracted from the set of 144 descriptors load most strongly.[17] Specifically, each table shows the ten descriptors with the largest positive factor loadings and the ten descriptors with the largest negative factor loadings. As discussed in Section 4.1, the factor loadings can be interpreted as regression coefficients that would be produced from regressing each factor on the descriptors. Observing similarities between the descriptors that have the largest positive and negative factor loadings for a given factor can be used to interpret the factors (Fabrigar et al., 2009). Each table shows the O*NET descriptor name, the task measure group(s) into which the descriptor was placed for the structured analysis described in Section 4.1, and the factor loading value.

The first factor (see Table 1) and the fourth factor (see Table 4) both load strongly and positively on O*NET descriptors that we labeled as NRI, indicating non-routine interactive tasks. After looking through the strongest descriptors associated with each factor, we determined that the first factor appears to capture tasks that align more strongly with non-routine tasks (e.g., Guiding, Directing, and Motivating Subordinates; Coordinating the Work and Activities of Others; Developing and Building Teams; Staffing Organizational Units) and the fourth factor appears to capture tasks that align more closely with interpersonal/people tasks (e.g., Self Control, Concern for Others, Deal With Unpleasant or Angry People, Social Orientation). Thus, we interpret the first factor as non-routine tasks and the fourth factor as interpersonal tasks.

[17] Similar interpretations exist for the factors extracted from the set of 109 descriptors, but only the 144 set is shown here for sake of brevity.

The second factor (see Table 2) loads most strongly on O*NET descriptors that appear to be related to cognitive or analytical tasks (e.g., Technology Design, Information Ordering, Fluency of Ideas, Inductive Reasoning, Flexibility of Closure)—nearly every one of the strongest positive descriptors was given an analytical or cognitive interpretation in our structured approach. Moreover, seven of the ten strongest positive descriptors belong to the O*NET Content Model descriptor grouping referenced as Cognitive Abilities (i.e., the first 3-digits of their O*NET Element ID are 1A1). Thus, we interpret this latent factor as analytical/cognitive tasks.

The third factor (see Table 3) has strong positive loadings for descriptors that mostly appear to be related to manual tasks (e.g., Response Orientation, Multilimb Coordination, Reaction Time, Performing General Physical Activities), so we interpret this latent factor as manual tasks. Finally, the fifth factor (see Table 5) loads heavily on what we determined to be routine tasks (e.g., Importance of Repeating Same Tasks, Degree of Automation, Processing Information), so we interpret the fifth factor as routine tasks.

While our exact interpretation of the factors from the unstructured approach can be debated, we are encouraged by the fact that each factor from a completely data-driven approach does seem to load strongly on related tasks. We are also encouraged to find that the interpretations from the unstructured approach (Interpersonal, Analytical/cognitive, Manual, Non-routine, Routine) were related to the task groups used in the structured approach (non-routine interactive, non-routine analytical, non-routine manual, routine manual, routine cognitive). This gives us more confidence in the quality of the interpretations as well as the credibility of the five task groups from Autor, Levy, and Murnane (2003) used in the structured approach.

5.2 Comparing Task-based Approaches to SOC-based Approach

Table 6 shows the grand average occupation distance in tasks for the quantitative task-based and SOC-based approaches that use 144 (109) O*NET descriptors in the top (bottom) panel. The main takeaway from Table 6 is that our quantitative task-based approaches to constructing occupational categories have lower grand average variance than the SOC-based approach in all but one calculation (Structured factor analysis, Method 2-144; 4 percent increase) wherein the task-based approach generates fewer occupational categories than the SOC groupings. The unstructured task-based approach always outperforms the SOC-based approach, reducing grand average occupation distance in tasks by about 8%. This high-level summary evidence suggests that our quantitative task-based approaches may be better than the SOC-based approach at grouping detailed occupations into categories with similar task and skill requirements.[18]

Next, we take a closer look by examining group-level measures. Tables 7 and 8 shows these results for our quantitative task-based structured and unstructured approaches, respectively, and Table 9 shows results for the SOC-based approach. First, the task-based approaches yield different groupings of detailed occupations than the SOC-based approach (see the SOC groups column in Tables 7 and 8). Second, the structured approach tends to cluster many detailed occupations into few occupational categories. Appendix Figures A1-A3 show

[18] We also calculated grand average occupation distance in factor scores, these results are presented in Appendix Table A7. We find that our task-based approaches outperform the SOC-based approach here as well.

histograms of the average within-occupation-group distance in tasks for the structured, unstructured, and SOC-based approaches, respectively. The unstructured approach tends to have a tighter distribution of within-occupation-group average distance than either the structured approach or SOC-based approach. Thus, we find that while our structured approach has a foundation in economic theory and in the literature, it has an undesirable outcome in practice of very lumpy occupational categories. Our unstructured approach seems to capture the essence of the structured approach while producing more evenly distributed categories.

We also compare approaches by focusing in on a few selected task measures. Table 10 shows average within-occupation-group variance in tasks for five separate O*NET task descriptors. We selected these tasks because they align well with the five latent skill factors: Coordinating the Work and Activities of Others (Factor 1, non-routine), Information Ordering (Factor 2, analytical/cognitive), Performing General Physical Activities (Factor 3, manual), Social Orientation (Factor 4, interpersonal), and Importance of Repeating Same Tasks (Factor 5, routine). For these individual tasks, we see that task-based approaches often produce a smaller average variance in tasks than the SOC-based groups. Moreover, the unstructured task-based approach using 144 descriptors always yields a smaller average variance in tasks than the SOC-based approach. To dispel any concerns that we might have "cherry-picked" the O*NET descriptors in Table 10, we present similar tables for the top 5 O*NET descriptors for each of the five latent skill factors in Appendix Tables A2-A6. The task-based approaches almost always have lower average within-occupation-group variance than the SOC-based approach.

6 Applications

6.1 Motivating Example: Teachers and Homebuilders

In the introduction we discussed that, in Hirsch and Manzella (2015), both teachers and homebuilders rank highly in “developing/teaching others,” which is one type of “caring.” We were interested in seeing how similar these occupations are in terms of latent skill/task factors generated from the task-based approach.

Based on the unstructured approach with 144 descriptors, the most similar latent skill/task factors related to the “developing/teaching others” trait are the first factor, shown in Table 1, and the fourth factor, shown in Table 4. We interpreted the first factor as measuring non-routine tasks. It loads strongly on descriptors such as Guiding, Directing, and Motivating Subordinates; Coordinating the Work and Activities of Others; Developing and Building Teams; and Coaching and Developing Others. We interpreted the fourth factor as measuring interpersonal tasks. It loads strongly on descriptors such as Concern for Others; Social Orientation; Assisting and Caring for Others; and Contact with Others.

The teacher occupations and the Manufactured Building and Mobile Home Installers occupation ended up in the same high-low group for both the non-routine skill factor and the interactive skill factor; both were grouped as higher than the median in each one. Overall, the teacher-related occupations all are classified into the occupational category having “high non-routine skills, high analytical/cognitive skills, low manual skills, low routine skills, high interactive skills,”, whereas the homebuilders occupation is classified into the occupational category having “high non-routine skills, low cognitive/analytical skills, high manual skills, low routine skills, high interactive skills.” Thus, the task-based occupation groups for these two occupations are very intuitive and also capture the nuanced similarities discussed in the introduction; the two occupations are similar in terms of their non-routine, interactive, and routine skill requirements, but teaching requires more cognitive/analytical skills while homebuilding requires more manual skills.

The full set of detailed occupations in each of these two task-based groups from the unstructured approach are shown in Tables 11 and 12. Another notable result in these tables is that the occupations within each group appear intuitive and yet they come from different SOC groups. For example, in Table 11, we see that teachers, librarians, speech-language pathologists, public relations specialists, coaches, actors, and agents all fall into the same task-based group even though they come from eight different two-digit SOC groups.

6.2 Illustrative Example: Police, Fire and Ambulance Dispatchers

In the background section, we discussed a recent example from the 2018 SOC revision process concerning whether public safety dispatchers perform tasks that are more similar to Protective Service occupations (SOC group 33-0000) than to Office and Administrative Support occupations (SOC group 43-0000). For the purposes of this research, we are interested in examining how our task-based approach groups these detailed occupations.

The detailed occupation of interest, Police, Fire, and Ambulance Dispatchers (43-5031), and another detailed occupation, Dispatchers, Except Police, Fire, and Ambulance (43-5032), are collapsed into the broad occupation, Dispatchers (43-4030) in the 2002 vintage of Census Occupation codes. Therefore, we build an O*NET dataset at the detailed occupation level based on the 2000 SOC six-digit codes instead of at the 2002 Census Occupation Code level; these details are discussed in the Data Appendix B. Then we construct occupational categories derived from our (preferred) task-based unstructured factor analysis approach and the O*NET data at the SOC-level with 144 descriptors.

Examining our task-based occupation categories, we see that most of the detailed occupations that are grouped into the SOC-based category Office and Administrative Support (43-0000) are also grouped together in one task-based category (referenced as grouping 29 in Table 13) that is characterized by the task composition of low non-routine, low cognitive, low manual, high interpersonal, and high routine. We also see that first-responder occupations (e.g., Police Officers, Fire Fighters) are mainly found in two groups (referenced as groupings 2 and 10 in Table 13) that are characterized by similar task compositions in four of the five factors—they differ in the relative amount of cognitive tasks. Overall, the task-based occupation groupings are intuitive.

Police, Fire, and Ambulance Dispatchers (43-5031) is not categorized into any of these aforementioned task-based occupation groups. Instead they are grouped into a category (referenced as grouping 13 in Table 13) that is characterized by the task composition of high non-routine, low cognitive, low manual, high interpersonal, and high routine. Police, Fire, and Ambulance Dispatchers are similar in most dimensions to most office/administrative support occupations (in four out of five task factors); they are also similar in many dimensions to first-responder occupations like Police Patrol Officers (in three out of five task factors) and Fire Fighters (in two out of five task factors). Recall that Factor 1 (non-routine tasks) explains most (47%) of the variation in tasks while Factor 5 (routine tasks) explains the least (2.5%). So it seems more relevant that Police, Fire, and Ambulance Dispatchers is aligned with the occupation group with first responders like Police Patrol Officers because both groups are categorized similarly along the first latent task factor. It seems less relevant that Police, Fire, and Ambulance Dispatchers is aligned with the occupation group with office/administrative support occupations because both groups are categorized similarly along the fifth latent task factor.

Reflecting back to arguments by respondents in the 2018 SOC revision process, the nature of work performed by Police, Fire, and Ambulance Dispatchers is described in ways that closely align with the O*NET descriptors having the strongest positive loadings in Factor 1 (non-routine), see Table 1. Task-specific examples given by respondents include the following:

- “[being] able to answer and prioritize multiple emergent and non-emergent telephone lines, send fire and/or medical responders, and dispatch law enforcement officers—under very specific set of policy guidelines”
- “[having a high amount of] responsibility of making split-second decisions in a time critical, error-free environment”
- “gathering and providing information to ensure the safest response to the incident”
- “questions the caller, selects and appropriate method (and level of response, provides pertinent information to responders (fire, medical, and law enforcement personnel) and gives appropriate aid and direction for patients through the caller”

Such tasks can be captured by the following set of O*NET descriptors: Coordinating the Work Activities of Others; Developing and Building Teams; Coaching and Developing Others; Scheduling Work and Activities; and Responsibility for Outcomes and Results. Thus, our task-based approach suggests that the similarity of tasks combinations performed by public safety dispatchers and first-responders is greater than the similarity of task combinations performed by public safety dispatchers and office/administrative support workers.

The above examples illustrates how our task-based approach can be used to provide more nuanced interpretation for grouping together detailed occupations that perform similar tasks. They also highlight that our task-based approach has the additional benefit of enabling quantitative assessments of similarities in particular skill/task bundles across occupations. Lastly, this example underscores that the SOC revision process may result in occupational categories that may not align well with researchers' needs.

6.3 Empirical Labor Economics: The Role of Occupations in the Gender Wage Gap

In addition to providing a quantitative approach to avoid potential subjective disagreements as discussed with the example in 6.2, our task-based occupational categories have potential applications for economic research. One such example involves the recent work by Blau and Khan (2017). The authors provide a review of the long literature on the magnitude and causes of the gender wage gap. They also provide modern estimates using the Panel Study of Income Dynamics (PSID) over the 1980-2010 period. They consider many potential contributors to the gender wage gap and find that differences in occupations between women and men accounts for a larger fraction of the wage gap than any other measurable characteristic. Specifically, they find that occupation differences can explain 33 percent of the 2010 gender wage gap (page 799, Table 4). We are interested in examining whether and to what extent the use of alternative occupational categories in this content may lead to different empirical results and possibly a deeper understanding in the analysis of labor market outcomes.

We replicate the analysis in Blau and Kahn (2017) first using the exact occupational categories in their analysis, which are predominantly two-digit SOC groups with a few changes.[19] Then we replicate their analysis using the unaltered two-digit SOC groups and our unstructured task-based groups. We use microdata from the U.S. Census Bureau's SIPP GSF focusing on workers in the 2008 SIPP panel. The GSF includes SIPP household survey data linked with administrative records on tax and benefit data from the Internal Revenue Service (IRS) and Social Security Administration (SSA).[20] The SIPP survey data includes rich demographic information including detailed occupation. We restrict our attention to estimating only the "full specification" discussed in Blau and Kahn's analysis (2017, page 797), which is a Mincerian wage regression model augmented with a series of occupation, industry, and unionization dummy variables.

Table 14 presents estimates from the decomposition of the gender wage gap using the SIPP GSF (2008 panel) where results are derived from survey reported earnings. Our estimated gender wage gap is 0.2697 log points, meaning that on average women earn roughly 27% less than men. Our estimate aligns well with Blau and Kahn's estimate of 0.2314 log points for 2010 using PSID data. We find that our task-based occupational categories explain more of variation in the gender gap than alternative SOC-based approaches explain. The SOC-based occupational categories explain about 6% (1.6 log points) of the gap while the task-based occupational categories explain about 9.6% (2.6 log points) of the gap.

There are some differences between our replication results and those in the Blau and Kahn (2017) study such as the total amount of variation explained and the importance of

[19] We provide a detailed discussion in Appendix C: Data Appendix for Replication.

[20] We do not show results derived from administrative records data in this paper due to time constraints. We previously presented such results at conferences and we may include such results in future drafts. Notably, the administrative records data provides the highest quality data on earnings (Abowd and Stinson, 2013; Meyer et al., 2015; Chenevert et al., 2016).

experience, industry and occupation variables. However, such differences are consistent across the three sets of occupational categories and may be due to differences across data sources. In subsequent analyses, we will estimate the decomposition of the gender wage gap using the PSID data from Blau and Kahn (2017) and the task-based and SOC-based occupational categories. Taken together, the SIPP GSF results and PSID results will shed light on how task-based alternative occupation groups may provide new insights into many economic questions.

7 Conclusion

There is a large literature that accounts for the role of occupations in the analysis of labor market outcomes by controlling for occupational categories, and these categories are often constructed using two-digit SOC codes. The main purpose of this study is assessing the potential importance of using a quantitative task-based approach to constructing occupational categories. In our approach, we utilize factor analysis with a large, rich set of O*NET job descriptors that provide continuous measures of tasks and skills required across all occupations. We introduce an unstructured approach akin to exploratory factor analysis and a structured approach that arranges O*NET descriptors into task groups, based on work by Autor, Levy, and Murnane (2003) that looks at trends in job skill demands, before performing factor analysis.

We find that our unstructured approach yields five interpretable latent skill factors (analytic/cognitive, non-routine, manual, interpersonal, routine) that align well with the theory behind the five latent skill factors in our structured approach (non-routine analytical, non-routine manual, routine manual, non-routine interactive, routine cognitive).

When we compare approaches, we show that the grand average occupation distance in tasks is lower in nearly all calculations for our task-based approaches. We also find that our structured approach has an undesirable outcome in practice of very lumpy occupational categories, whereas our unstructured approach seems to capture the essence of the structured approach while producing more evenly distributed categories. Our unstructured approach also has a tighter distribution of within-occupation-group occupation distance in tasks than either our structured approach or the SOC-based approach. We also examine a few selected individual O*NET descriptors that seem characteristic of the five latent skill factors in our unstructured approach. We find that both the unstructured and structured task-based approaches produce a smaller average variance in tasks than the SOC-based approach. This is encouraging evidence that our unstructured task-based approach does a better job of grouping together occupations with similar task/skill requirements than the SOC-based approach.

When we return to our motivating example of teachers and homebuilders, indeed we see that these two occupations are similar in terms of their non-routine, interpersonal, and routine skill requirements; however, teaching requires more analytical/cognitive skills while homebuilding requires more manual skills. This application also provides a sense that the detailed occupations are reasonably grouped into occupational categories using our unstructured task-based approach even though these detailed occupations belong to different occupational categories based on two-digit SOC codes.

Additionally, we examine a recent example from the SOC revision process concerning whether the tasks performed by Police, Fire, and Ambulance Dispatchers are more closely aligned with the tasks performed by Office and Administrative Support Occupations or the tasks performed by Protective Service Occupations. Our task-based approach suggests that the task bundle performed by public safety dispatchers is better aligned with the task bundle performed by first responders like Police Officers than with the task bundle performed by most office and administrative support workers.

Lastly, we examine whether our task-based occupational categories do a better job of explaining occupational differences in earnings by replicating a recent analysis by Blau and Kahn (2017) on the gender wage gap. We find our task-based occupational categories explain more of the gender wage gap than the SOC-based approaches explain.

Taken together, these applications highlight many important contributions of using our task-based approach such as its intuitiveness, interpretability and imbedded nuances, its quantitative assessment capabilities, and its resulting occupation groups that may better serve researchers' needs.

In future work, we focus our attention on some next steps. We will estimate the decomposition of the gender wage gap using the PSID data from Blau and Kahn (2017) and the task-based and SOC-based occupational categories. We would also like to examine flows of workers within and across occupation groups in order to better understand differences between the task-based and SOC-based occupational categories. Additionally, we plan to examine whether, and to what extent, using a Mahalanobis occupation distance measure instead of an Euclidean distance measure matters. Task variables are strongly correlated. Using a Mahalanobis occupation distance measure would get rid of any potential scaling effects or collinearity effects of the task variables. While we do not expect the main takeaways to change when replacing our Euclidean distance results with Mahalanobis distance results, documenting any differences between these occupation distance results is an additional contribution to the literature on human capital, task-specificity, and occupational mobility.

References

Abowd, J. M., & Stinson, M. H. (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics*, *95*(5), 1451-1467.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, *118*(4), 1279-1333.

Bartlett, M.S. (1937). The statistical conception of mental factors. *British Journal of Psychology, 28,* 97-104.

Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, *55*(3), 789-865.

Chenevert, R., M. Klee, and K. Wilkin (2016). Do imputed earnings earn their keep? Evaluating SIPP earnings and nonresponse with administrative records. U.S. Census Bureau Working Paper Number: SEHSD-WP2016-18, SIPP-WP-275.

Dey, M., & M. A. Lowenstien (2019) On Job Requirements, Skill, and Wages. U.S. Bureau of Labor Statistics Working Paper Number: 513.

Fabrigar, L.R., MacCallum, R.C., Wegener, D.T., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4(3),* 272-299.

Hirsch, B. T. (2004). Reconsidering union wage effects: Surveying new evidence on an old topic. *Journal of Labor Research*, *25*(2), 233-266.

Hirsch, B. T., & Manzella, J. (2015). Who Cares–and Does it Matter? Measuring Wage Penalties for Caring Work. *Research In Labor Economics*, 41, 213-275.

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199-226.

Occupational Information Network O*NET https://www.onetcenter.org/database.html

O'Neill, J. (1990). The role of human capital in earnings differences between black and white men. *Journal of Economic Perspectives*, *4*(4), 25-45.

Tables and Figures

Table 1. Strongest and weakest loadings for Factor 1 (unstructured task-based approach, 144 O*NET descriptors, correlated factors)

	O*NET descriptor name	O*NET Element ID	O*NET Content Model 3-digit reference group	Structured Approach Task Measure Group	Factor Loading
Strongest Positive Descriptors	Guiding, Directing, and Motivating Subordinates	4A4b4	Interacting With Others	NRI	0.9931
	Coordinating the Work and Activities of Others	4A4b1	Interacting With Others	NRI	0.9362
	Developing and Building Teams	4A4b2	Interacting With Others	NRI	0.9105
	Staffing Organizational Units	4A4c2	Interacting With Others	NRI	0.8989
	Coaching and Developing Others	4A4b5	Interacting With Others	NRI	0.8440
	Monitoring and Controlling Resources	4A4c3	Interacting With Others	NRA, NRI	0.8200
	Scheduling Work and Activities	4A2b5	Mental Processes	NRA, NRI, NRM	0.7594
	Management of Material Resources	2B5c	Resource Management Skills	NRM	0.7455
	Responsibility for Outcomes and Results	4C1c2	Interpersonal Relationships	NRI	0.7197
	Management of Financial Resources	2B5b	Resource Management Skills	NRA, NRI	0.7065
Strongest Negative Descriptors	Independence	1C6	Independence	NRA, NRI, NRM	-0.1111
	Deal With Unpleasant or Angry People	4C1d2	Interpersonal Relationships	NRI	-0.1193
	Installation	2B3d	Technical Skills	NRM	-0.1505
	Degree of Automation	4C3b2	Structural Job Characteristics	RC, RM	-0.1577
	Wrist-Finger Speed	1A2c2	Psychomotor Abilities	RM	-0.1617
	Control Precision	1A2b1	Psychomotor Abilities	RM	-0.1732
	Manual Dexterity	1A2a2	Psychomotor Abilities	NRM, RM	-0.2723
	Finger Dexterity	1A2a3	Psychomotor Abilities	NRM, RM	-0.2737
	Arm-Hand Steadiness	1A2a1	Psychomotor Abilities	NRM, RM	-0.3154
	Importance of Repeating Same Tasks	4C3b7	Structural Job Characteristics	RC, RM	-0.3460
Interpretation	*Non-routine*				

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024.

The five structured approach task measure groups are denoted as non-routine interactive tasks (NRI), non-routine analytical tasks (NRA), routine cognitive tasks (RC), routine manual tasks (RM), and non-routine manual tasks (NRM).

Table 2. Strongest and weakest loadings for Factor 2 (unstructured task-based approach, 144 O*NET descriptors, correlated factors)

	O*NET descriptor name	O*NET Element ID	O*NET Content Model 3-digit reference group	Structured Approach Task Measure Group	Factor Loading
Strongest Positive Descriptors	Technology Design	2B3b	Technical Skills	NRA	0.6752
	Information Ordering	1A1b6	Cognitive Abilities	RC	0.6725
	Inductive Reasoning	1A1b5	Cognitive Abilities	NRA	0.6575
	Flexibility of Closure	1A1e2	Cognitive Abilities	RC	0.6506
	Category Flexibility	1A1b7	Cognitive Abilities	RC	0.6465
	Fluency of Ideas	1A1b1	Cognitive Abilities	NRA	0.6430
	Originality	1A1b2	Cognitive Abilities	NRA	0.6402
	Speed of Closure	1A1e1	Cognitive Abilities	NRA	0.6339
	Active Learning	2A2b	Process	NRA, NRI, NRM, RC, RM	0.6175
	Updating and Using Relevant Knowledge	4A2b3	Mental Processes	NRM	0.6122
Strongest Negative Descriptors	Trunk Strength	1A3a4	Physical Abilities	NRM, RM	-0.2018
	Rate Control	1A2b4	Psychomotor Abilities	NRM	-0.2022
	Stamina	1A3b1	Physical Abilities	NRM, RM	-0.2230
	Frequency of Conflict Situations	4C1d1	Interpersonal Relationships	NRI	-0.2238
	Speed of Limb Movement	1A2c3	Psychomotor Abilities	NRM, RM	-0.2536
	Deal With Unpleasant or Angry People	4C1d2	Interpersonal Relationships	NRI	-0.3017
	Degree of Automation	4C3b2	Structural Job Characteristics	RC, RM	-0.3036
	Responsibility for Outcomes and Results	4C1c2	Interpersonal Relationships	NRI	-0.3379
	Responsible for Others' Health and Safety	4C1c1	Interpersonal Relationships	NRI	-0.4240
	Pace Determined by Speed of Equipment	4C3d3	Structural Job Characteristics	NRM, RM	-0.4596
Interpretation	*Analytical/Cognitive*				

Data source: O*NET database version 15.0. Census DRB release CBDRB-FY19-CED001-B0024.

The five structured approach task measure groups are denoted as non-routine interactive tasks (NRI), non-routine analytical tasks (NRA), routine cognitive tasks (RC), routine manual tasks (RM), and non-routine manual tasks (NRM).

Table 3. Strongest and weakest loadings for Factor 3 (unstructured task-based approach, 144 O*NET descriptors, correlated factors)

	O*NET descriptor name	O*NET Element ID	O*NET Content Model 3-digit reference group	Structured Approach Task Measure Group	Factor Loading
Strongest Positive Descriptors	Response Orientation	1A2b3	Psychomotor Abilities	NRM	0.9023
	Multilimb Coordination	1A2b2	Psychomotor Abilities	NRM, RM	0.8986
	Reaction Time	1A2c1	Psychomotor Abilities	NRM	0.8713
	Operation and Control	2B3h	Technical Skills	NRM, RM	0.8696
	Performing General Physical Activities	4A3a1	Work Output	RM	0.8677
	Control Precision	1A2b1	Psychomotor Abilities	RM	0.8627
	Troubleshooting	2B3k	Technical Skills	NRA	0.8376
	Static Strength	1A3a1	Physical Abilities	NRM, RM	0.8374
	Operation Monitoring	2B3g	Technical Skills	RC	0.8329
	Repairing and Maintaining Mechanical Equipment	4A3b4	Work Output	NRM	0.8326
Strongest Negative Descriptors	Written Comprehension	1A1a2	Cognitive Abilities	NRI	-0.3222
	Interacting With Computers	4A3b1	Work Output	NRA	-0.3262
	Writing	2A1c	Content	NRI	-0.3396
	Speech Clarity	1A4b5	Sensory Abilities	NRI	-0.3562
	Letters and Memos	4C1a2j	Interpersonal Relationships	NRI	-0.3592
	Written Expression	1A1a4	Cognitive Abilities	NRI	-0.3593
	Speech Recognition	1A4b4	Sensory Abilities	NRI	-0.3696
	Active Listening	2A1b	Content	NRI	-0.3843
	Speaking	2A1d	Content	NRI	-0.3872
	Electronic Mail	4C1a2h	Interpersonal Relationships	NRI	-0.4775
Interpretation	*Manual*				

Data source: O*NET database version 15.0. Census DRB release CBDRB-FY19-CED001-B0024.
The five structured approach task measure groups are denoted as non-routine interactive tasks (NRI), non-routine analytical tasks (NRA), routine cognitive tasks (RC), routine manual tasks (RM), and non-routine manual tasks (NRM).

Table 4. Strongest and weakest loadings for Factor 4 (unstructured task-based approach, 144 O*NET descriptors, correlated factors)

	O*NET descriptor name	O*NET Element ID	O*NET Content Model 3-digit reference group	Structured Approach Task Measure Group	Factor Loading
Strongest Positive Descriptors	Self Control	1C4a	Adjustment	NRI	0.8252
	Concern for Others	1C3b	Interpersonal Orientation	NRI	0.8197
	Deal With Unpleasant or Angry People	4C1d2	Interpersonal Relationships	NRI	0.7461
	Social Orientation	1C3c	Interpersonal Orientation	NRI	0.7358
	Assisting and Caring for Others	4A4a5	Interacting With Others	NRI	0.7136
	Contact With Others	4C1a4	Interpersonal Relationships	NRI	0.7075
	Performing for or Working Directly with the Public	4A4a8	Interacting With Others	NRI	0.6925
	Deal With Physically Aggressive People	4C1d3	Interpersonal Relationships	NRI	0.6851
	Deal With External Customers	4C1b1f	Interpersonal Relationships	NRI	0.6834
	Frequency of Conflict Situations	4C1d1	Interpersonal Relationships	NRI	0.6232
Strongest Negative Descriptors	Equipment Maintenance	2B3j	Technical Skills	NRM	-0.2335
	Visualization	1A1f2	Cognitive Abilities	NRA	-0.2354
	Programming	2B3e	Technical Skills	NRA	-0.2382
	Troubleshooting	2B3k	Technical Skills	NRA	-0.2477
	Technology Design	2B3b	Technical Skills	NRA	-0.2595
	Quality Control Analysis	2B3m	Technical Skills	NRA	-0.2888
	Pace Determined by Speed of Equipment	4C3d3	Structural Job Characteristics	NRM, RM	-0.2958
	Equipment Selection	2B3c	Technical Skills	NRM	-0.3027
	Estimating the Quantifiable Characteristics of Products, Events, or Information	4A1b3	Information Input	NRA, NRM	-0.3050
	Drafting, Laying Out, and Specifying Technical Devices, Parts, and Equipment	4A3b2	Structural Job Characteristics	NRM	-0.3888
Interpretation	*Interpersonal*				

Data source: O*NET database version 15.0. Census DRB release CBDRB-FY19-CED001-B0024.
The five structured approach task measure groups are denoted as non-routine interactive tasks (NRI), non-routine analytical tasks (NRA), routine cognitive tasks (RC), routine manual tasks (RM), and non-routine manual tasks (NRM).

Table 5. Strongest and weakest loadings for Factor 5 (unstructured task-based approach, 144 O*NET descriptors, correlated factors)

	O*NET descriptor name	O*NET Element ID	O*NET Content Model 3-digit reference group	Structured Approach Task Measure Group	Factor Loading
Strongest Positive Descriptors	Importance of Repeating Same Tasks	4C3b7	Structural Job Characteristics	RC, RM	0.6742
	Degree of Automation	4C3b2	Structural Job Characteristics	RC, RM	0.6724
	Consequence of Error	4C3a1	Structural Job Characteristics	NRA, NRI, NRM	0.4785
	Processing Information	4A2a2	Mental Processes	RC	0.4757
	Evaluating Information to Determine Compliance with Standards	4A2a3	Mental Processes	NRA, NRM	0.4683
	Perceptual Speed	1A1e3	Cognitive Abilities	RC	0.4593
	Documenting/Recording Information	4A3b6	Work Output	RC	0.4348
	Selective Attention	1A1g1	Cognitive Abilities	NRA, NRI, NRM, RC, RM	0.4027
	Interacting With Computers	4A3b1	Work Output	NRA	0.3876
	Monitor Processes, Materials, or Surroundings	4A1a2	Information Input	RC	0.3622
Strongest Negative Descriptors	Performing General Physical Activities	4A3a1	Work Output	RM	-0.2141
	Extent Flexibility	1A3c1	Physical Abilities	NRM, RM	-0.2302
	Selling or Influencing Others	4A4a6	Interacting With Others	NRI	-0.2465
	Performing for or Working Directly with the Public	4A4a8	Interacting With Others	NRI	-0.2566
	Innovation	1C7a	Practical Intelligence	NRA, NRI, NRM	-0.2651
	Gross Body Coordination	1A3c3	Physical Abilities	NRM, RM	-0.2743
	Stamina	1A3b1	Physical Abilities	NRM, RM	-0.2788
	Dynamic Strength	1A3a3	Physical Abilities	NRM, RM	-0.2826
	Trunk Strength	1A3a4	Physical Abilities	NRM, RM	-0.3223
	Dynamic Flexibility	1A3c2	Physical Abilities	NRM, RM	-0.3717
Interpretation	*Routine*				

Data source: O*NET database version 15.0. Census DRB release CBDRB-FY19-CED001-B0024.
The five structured approach task measure groups are denoted as non-routine interactive tasks (NRI), non-routine analytical tasks (NRA), routine cognitive tasks (RC), routine manual tasks (RM), and non-routine manual tasks (NRM).

Table 6. Average occupation distance in tasks across all detailed occupations

Top panel: 144 O*NET descriptors			
	SOC-based approach	Task-based approaches	
		Structured factor-analysis	Unstructured factor-analysis
average occupation distance	1.122706	1.162799	1.042798
number of occupational categories	22	19	31
Bottom panel: 109 O*NET descriptors			
	SOC-based approach	Task-based approaches	
		Structured factor-analysis	Unstructured factor-analysis
average occupation distance	0.983212	0.965001	0.899403
number of occupational categories	22	25	31

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024

Occupation distance measure A is the Euclidean distance in tasks (i.e., mean scaled O*NET descriptor rating values) between a detailed occupation and the mean of the occupational category. Average occupation distance in tasks across all occupations is calculated by (1) calculating occupation distance measure A for each detailed occupation, and (2) calculating the average distance in step 1 across all 485 detailed occupations.

Table 7. Average within-occupation-group distance in tasks, structured task-based approach

	Occupational category construction					Task-based Approach, Structured factor analysis					
	NRI	NRA	NRM	RM	RC	Method 1-109			Method 2-144		
						distance	N	SOC groups	distance	N	SOC groups
1	hi	hi	hi	hi	hi	1.030472	42	10	1.191077	43	11
2	hi	hi	hi	hi	lo	0.803215	4	4	1.044087	4	4
3	hi	hi	hi	lo	hi	0.979347	26	9	--	0	--
4	hi	hi	hi	lo	lo	--	0	--	0	1	1
5	hi	hi	lo	hi	hi	0.785855	11	4	1.007640	14	7
6	hi	hi	lo	hi	lo	0.737561	2	2	0	1	1
7	hi	hi	lo	lo	hi	1.062910	112	16	1.258538	159	17
8	hi	hi	lo	lo	lo	0.997658	20	9	0.963827	6	6
9	hi	lo	hi	hi	hi	0	0	--	--	0	--
10	hi	lo	hi	hi	lo	0	1	1	0	1	1
11	hi	lo	hi	lo	hi	--	1	1	--	0	--
12	hi	lo	hi	lo	lo	0	0	--	--	0	--
13	hi	lo	lo	hi	hi	0	1	1	--	0	--
14	hi	lo	lo	hi	lo	0	1	1	--	0	--
15	hi	lo	lo	lo	hi	0.912773	5	3	0.789194	3	1
16	hi	lo	lo	lo	lo	1.030057	16	6	0.997587	10	4
17	lo	hi	hi	hi	hi	0.951465	13	5	0.944911	8	3
18	lo	hi	hi	hi	lo	0.754195	6	3	0.793535	3	2
19	lo	hi	hi	lo	hi	0	1	1	--	0	--
20	lo	hi	hi	lo	lo	--	0	--	0	1	1
21	lo	hi	lo	hi	hi	--	0	--	--	0	--
22	lo	hi	lo	hi	lo	0	1	1	--	0	--
23	lo	hi	lo	lo	hi	0.816536	3	3	0.884662	2	2
24	lo	hi	lo	lo	lo	0	1	1	--	0	--
25	lo	lo	hi	hi	hi	0.862864	20	4	1.038379	8	6
26	lo	lo	hi	hi	lo	0.940595	124	11	1.156251	160	11
27	lo	lo	hi	lo	hi	--	0	--	--	0	--
28	lo	lo	hi	lo	lo	0.887903	4	4	1.149667	13	8
29	lo	lo	lo	hi	hi	--	0	--	--	0	--
30	lo	lo	lo	hi	lo	0.944987	16	7	--	0	--
31	lo	lo	lo	lo	hi	0.783345	7	3	0.872270	5	3
32	lo	lo	lo	lo	lo	1.032272	47	11	1.206990	43	8

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024
N denotes the number of detailed occupations in a task-based category. SOC groups refers to the number of two-digit SOC categories within a task-based category. Occupational category numbering does not imply any correspondence across approaches. That is, the occupational category 1 constructed by approach A may contain a different set of detailed occupations than the occupational category 1 constructed by approach B. The median factor score value is used as the high-low threshold for a given latent skill factor, see text for discussion. Occupation distance measure A is the Euclidean distance in tasks (i.e., mean scaled O*NET descriptor rating values) between a detailed occupation and the mean of the occupational category. Average within-occupation-group distance is calculated by (1) calculating occupation distance measure A for each detailed occupation, and (2) calculating the average distance in step 1 across the number of detailed occupations within the occupational category.

Table 8. Average within-occupation-group distance in tasks, unstructured task-based approach

	Occupational category construction					Task-based Approach, Unstructured factor analysis					
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Correlated-109			Correlated-144		
						distance	N	SOC groups	distance	N	SOC groups
1	hi	hi	hi	hi	hi	0.984766	35	10	1.130457	22	7
2	hi	hi	hi	hi	lo	1.046229	11	6	1.189634	15	8
3	hi	hi	hi	lo	hi	0.862108	8	5	0.972769	12	5
4	hi	hi	hi	lo	lo	0.976322	17	7	1.156204	4	4
5	hi	hi	lo	hi	hi	0.936013	27	9	1.103298	40	11
6	hi	hi	lo	hi	lo	0.880022	40	12	1.107637	31	7
7	hi	hi	lo	lo	hi	0.771886	4	2	1.068061	41	5
8	hi	hi	lo	lo	lo	0.980077	32	7	1.166636	8	6
9	hi	lo	hi	hi	hi	0.768754	4	2	1.088149	14	6
10	hi	lo	hi	hi	lo	0.829640	6	3	1.122317	9	6
11	hi	lo	hi	lo	hi	0.806050	6	3	1.032388	11	4
12	hi	lo	hi	lo	lo	0.906136	10	5	0.913431	21	6
13	hi	lo	lo	hi	hi	0.916025	5	4	0.958776	9	3
14	hi	lo	lo	hi	lo	0.829970	22	6	0.986341	4	3
15	hi	lo	lo	lo	hi	0	1	1	0	1	1
16	hi	lo	lo	lo	lo	0.944494	14	7	--	0	--
17	lo	hi	hi	hi	hi	0.850296	7	4	0.583298	3	1
18	lo	hi	hi	hi	lo	0.934757	9	5	1.019539	4	3
19	lo	hi	hi	lo	hi	0.788555	26	5	1.007628	8	4
20	lo	hi	hi	lo	lo	0.786357	8	3	1.011261	22	3
21	lo	hi	lo	hi	hi	0.919037	9	5	0.962826	12	8
22	lo	hi	lo	hi	lo	0.967706	7	4	1.128358	11	4
23	lo	hi	lo	lo	hi	--	0	--	1.026133	7	6
24	lo	hi	lo	lo	lo	0.837935	2	1	0.632799	2	2
25	lo	lo	hi	hi	hi	0.789376	7	3	1.005786	4	2
26	lo	lo	hi	hi	lo	0.844651	4	2	1.073369	15	8
27	lo	lo	hi	lo	hi	0.804844	59	5	0.898740	28	4
28	lo	lo	hi	lo	lo	0.914678	25	7	0.999478	50	8
29	lo	lo	lo	hi	hi	0.982955	24	9	0.966133	24	4
30	lo	lo	lo	hi	lo	0.938185	25	6	1.114685	25	7
31	lo	lo	lo	lo	hi	0.866640	20	6	1.152710	6	2
32	lo	lo	lo	lo	lo	1.117273	11	8	1.132741	22	10

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024
N denotes the number of detailed occupations in a task-based category. SOC groups refers to the number of two-digit SOC categories within a task-based category. Occupational category numbering does not imply any correspondence across approaches. That is, the occupational category 1 constructed by approach A may contain a different set of detailed occupations than the occupational category 1 constructed by approach B. The median factor score value is used as the high-low threshold for a given latent skill factor, see text for discussion. Occupation distance measure A is the Euclidean distance in tasks (i.e., mean scaled O*NET descriptor rating values) between a detailed occupation and the mean of the occupational category. Average within-occupation-group distance is calculated by (1) calculating occupation distance measure A for each detailed occupation, and (2) calculating the average distance in step 1 across the number of detailed occupations within the occupational category.

Table 9. Average within-occupation-group distance in tasks, SOC-based approach

Occupational category construction			SOC-based Approach		
2002 Census Code	2000 SOC Code	Title	N	SOC2000-109	SOC2000-144
0010-0430	11-0000	Management Occupations	27	1.019737	1.132331
0500-0950	13-0000	Business and Financial Operations Occupations	24	0.921511	1.048416
1000-1240	15-0000	Computer and Mathematical Occupations	11	0.974226	1.076632
1300-1560	17-0000	Architecture and Engineering Occupations	21	0.870083	0.973461
1600-1960	19-0000	Life, Physical, and Social Science Occupations	21	1.067215	1.181108
2000-2060	21-0000	Community and Social Service Occupations	5	0.696741	0.822417
2100-2160	23-0000	Legal Occupations	4	0.900498	0.985387
2200-2550	25-0000	Education, Training, and Library Occupations	11	0.853919	0.965891
2600-2960	27-0000	Arts, Design, Entertainment, Sports, and Media Occupations	17	1.002252	1.172882
3000-3540	29-0000	Healthcare Practitioners and Technical Occupations	28	0.917244	1.083290
3600-3650	31-0000	Healthcare Support Occupations	6	0.703899	0.799211
3700-3950	33-0000	Protective Service Occupations	16	1.146340	1.289615
4000-4160	35-0000	Food Preparation and Serving Related Occupations	12	0.914353	1.058664
4200-4250	37-0000	Building and Grounds Cleaning and Maintenance Occupations	6	1.195586	1.340665
4300-4650	39-0000	Personal Care and Service Occupations	19	1.057790	1.202975
4700-4960	41-0000	Sales and Related Occupations	16	1.096420	1.274847
5000-5930	43-0000	Office and Administrative Support Occupations	48	1.046030	1.204280
6005-6130	45-0000	Farming, Fishing, and Forestry Occupations	9	1.053461	1.214639
6200-6940	47-0000	Construction and Extraction Occupations	39	0.930582	1.064949
7000-7620	49-0000	Installation, Maintenance, and Repair Occupations	36	0.965021	1.091323
7700-8960	51-0000	Production Occupations	77	0.930455	1.072687
9000-9750	53-0000	Transportation and Material Moving Occupations	32	1.121519	1.286031

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024
N denotes the number of detailed occupations in an occupational category.
Occupation distance measure A is the Euclidean distance in tasks (i.e., mean scaled O*NET descriptor rating values) between a detailed occupation and the mean of the occupational category. Average within-occupation-group distance is calculated by (1) calculating occupation distance measure A for each detailed occupation, and (2) calculating the average distance in step 1 across the number of detailed occupations within the occupational category.

Table 10. Average variance in tasks for selected O*NET descriptors by approach

*O*NET task/skill descriptor*	Coordinating the Work and Activities of Others	Information Ordering	Performing General Physical Activities	Social Orientation	Importance of Repeating Same Tasks
*O*NET descriptor id*	4A4b1	1A1b6	4A3a1	1C3c	4C3b7
Latent skill factor	1	2	3	4	5
Factor Interpretation	Non-routine	Analytical/Cognitive	Manual	Interpersonal	Routine
SOC-based approaches					
SOC 2010	0.012791	0.002112	0.012215	0.010682	0.017932
Task-based approaches					
Structured factor-analysis					
Method 1-109	0.011235	0.001683	0.009240	0.014079	n/a
Method 2-144	0.011550	0.001836	0.012353	0.014929	0.021820
Unstructured factor-analysis					
Correlated-109	0.007920	0.001663	0.009483	0.009806	n/a
Correlated-144	0.008027	0.001538	0.010334	0.010656	0.015749

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024.
n/a indicates this O*NET descriptor was not included in the subset containing 109 variables.
Average variance in tasks is calculated by (1) calculating the variance of a selected task (i.e., mean scaled O*NET descriptor rating values) within each occupational category, and then (2) calculating the average of the variance in step 1 across all detailed occupations.

Table 11. Detailed occupations in the unstructured group having High non-routine skills, High cognitive/analytical skills, Low manual skills, Low routine skills, High interpersonal skills

Census 2002 Occupation Title	Census 2002 Code	SOC 2000 Code
Advertising and promotions managers	0040	11-2011
Marketing and sales managers	0050	11-2020
Public relations managers	0060	11-2031
Human resources managers	0130	11-3040
Education administrators	0230	11-9030
Social and community service managers	0420	11-9151
Agents and business managers of artists, performers, and athletes	0500	13-1011
Purchasing agents and buyers, farm products	0510	13-1021
Wholesale and retail buyers, except farm products	0520	13-1022
Meeting and convention planners	0720	13-1121
Counselors	2000	21-1010
Social workers	2010	21-1020
Clergy	2040	21-2011
Directors, religious activities and education	2050	21-2021
Postsecondary teachers	**2200**	**25-1000**
Preschool and kindergarten teachers	**2300**	**25-2010**
Elementary and middle school teachers	**2310**	**25-2020**
Secondary school teachers	**2320**	**25-2030**
Special education teachers	**2330**	**25-2040**
Other teachers and instructors	**2340**	**25-3000**
Librarians	2430	25-4021
Other education, training, and library workers	2550	25-90XX
Actors	2700	27-2011
Athletes, coaches, umpires, and related workers	2720	27-2020
Musicians, singers, and related workers	2750	27-2040
Public relations specialists	2820	27-3031
Speech-language pathologists	3230	29-1127
First-line supervisors/managers of retail sales workers	4700	41-1011
First-line supervisors/managers of non-retail sales workers	4710	41-1012
Travel agents	4830	41-3041
Real estate brokers and sales agents	4920	41-9020

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024

Table 12. Detailed occupations in the unstructured group having High non-routine skills, Low cognitive/analytical skills, High manual skills, Low routine skills, High interpersonal skills

Census 2002 Occupation Title	Census 2002 Code	SOC 2000 Code
Food service managers	0310	11-9051
Transit and railroad police	3860	33-3052
Bartenders	4040	35-3011
First-line supervisors/managers of landscaping, lawn service, and groundskeeping workers	4210	37-1012
Pest control workers	4240	37-2021
Septic tank servicers and sewer pipe cleaners	6750	47-4071
Miscellaneous construction and related workers	6760	47-4090
Automotive glass installers and repairers	7160	49-3022
Manufactured building and mobile home installers	**7550**	**49-9095**

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024

Table 13. Task-based Occupational Categories for the Illustrative Example of Dispatchers

	Occupational category construction					Task-based Approach, Unstructured factor analysis, Correlated 144				
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Detailed occupation of interest	N	SOC groups	SOC-33	SOC-43
1	hi	hi	hi	hi	hi		24	8	1	0
2	hi	hi	hi	hi	lo	Fire Fighters	31	12	6	0
3	hi	hi	hi	lo	hi		35	11	0	2
4	hi	hi	hi	lo	lo		8	6	0	0
5	hi	hi	lo	hi	hi		44	12	0	0
6	hi	hi	lo	hi	lo		63	10	0	1
7	hi	hi	lo	lo	hi		54	7	0	0
8	hi	hi	lo	lo	lo		28	5	0	0
9	hi	lo	hi	hi	hi		19	5	1	0
10	hi	lo	hi	hi	lo	Police Patrol Officers, Ambulance Drivers	13	5	4	0
11	hi	lo	hi	lo	hi		12	5	0	0
12	hi	lo	hi	lo	lo		3	1	0	0
13	hi	lo	lo	hi	hi	Police, Fire & Ambulance Dispatchers	20	9	1	3
14	hi	lo	lo	hi	lo		9	6	1	0
15	hi	lo	lo	lo	hi		5	5	0	1
16	hi	lo	lo	lo	lo		5	2	0	0
17	lo	hi	hi	hi	hi		6	5	0	0
18	lo	hi	hi	hi	lo		12	8	0	0
19	lo	hi	hi	lo	hi		13	3	0	1
20	lo	hi	hi	lo	lo		20	7	0	1
21	lo	hi	lo	hi	hi	Dispatchers, Except Police, Fire & Ambulance	14	8	0	3
22	lo	hi	lo	hi	lo		10	5	0	0
23	lo	hi	lo	lo	hi		5	3	0	0
24	lo	hi	lo	lo	lo		6	4	0	0
25	lo	lo	hi	hi	hi		13	6	1	1
26	lo	lo	hi	hi	lo		26	10	0	1
27	lo	lo	hi	lo	hi		58	5	0	0
28	lo	lo	hi	lo	lo		80	8	0	1
29	lo	lo	lo	hi	hi	Customer Service Representatives, Tellers, Secretaries	36	9	1	24
30	lo	lo	lo	hi	lo	Security Guards, Clerical Library Assistants	33	9	4	3
31	lo	lo	lo	lo	hi	Data Entry Keyers	15	8	0	5
32	lo	lo	lo	lo	lo	Postal Service Mail Carriers, File Clerks	26	10	0	5

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024.
Notes: O*NET data set is collapsed at SOC 2000 six-digit level and consists of 746 detailed occupations.
N denotes the number of detailed occupations in a task-based category. SOC groups refers to the number of two-digit SOC categories within a task-based category. SOC-33 refers to the number of detailed occupations that are in the two-digit SOC category, 33-0000, Protective Service Occupations. SOC-43 refers to the number of detailed occupations that are in the two-digit SOC category, 43-0000, Office and Administrative Support Occupations.

Table 14. Decomposition of the gender wage gap

Wages based on SIPP earnings from 2008 SIPP panel						
	Blau & Kahn occupational categories		SOC-based occupational categories		Task-based occupational categories Unstructured, Correlated-144	
	Log points	Percent of total gender gap	Log points	Percent of total gender gap	Log points	Percent of total gender gap
Total pay gap	0.2697	100.00	0.2697	100.00	0.2697	100.00
Total unexplained gap	0.1820	67.48	0.1928	71.49	0.1806	66.96
Total explained gap	0.0877	32.52	0.0768	28.48	0.0890	33.00
Total explained due to:						
Education variables	-0.0040	-1.48	-0.0039	-1.45	-0.0034	-1.26
Experience variables	0.0104	3.86	0.0105	3.89	0.0110	4.08
Region variables	0.0010	0.37	0.0010	0.37	0.0010	0.37
Race and ethnicity variables	0.0022	0.82	0.0022	0.82	0.0020	0.74
Unionization	0.0037	1.37	0.0035	1.30	0.0033	1.22
Industry variables	0.0501	18.58	0.0477	17.68	0.0500	18.54
Occupation variables	**0.0242**	**8.97**	**0.0158**	**5.86**	**0.0260**	**9.64**
Number of observations	25,000		25,000		25,000	

Data source: SIPP Gold Standard File (GSF). Census DRB release number CBDRB-FY19-501.
Sample includes individuals aged 25-64 at the time of their 2008 SIPP survey who have worked at least 26 weeks and have a wage value greater than or equal to $2/hour. Observations with missing data for variables of interest are dropped. No weights are used in these tables. See Appendix C for more details.

Appendix A: Tables and Figures

Appendix Table A1. Eigenvalues of top 20 latent skill/task factors derived from factor analysis, Unstructured task-based approach, Correlated-144

Factor	Eigenvalue	Difference	Proportion	Cumulative proportion
Factor 1	60.34705	36.89636	0.4670	0.4670
Factor 2	23.45068	13.10671	0.1815	0.6484
Factor 3	10.34397	6.02926	0.0800	0.7285
Factor 4	4.31471	1.14423	0.0334	0.7619
Factor 5	3.17049	0.20700	0.0245	0.7864
Factor 6	2.96349	0.13836	0.0229	0.8093
Factor 7	2.82513	0.64518	0.0219	0.8312
Factor 8	2.17995	0.26997	0.0169	0.8481
Factor 9	1.90998	0.27345	0.0148	0.8628
Factor 10	1.63653	0.20206	0.0127	0.8755
Factor 11	1.43447	0.11486	0.0111	0.8866
Factor 12	1.31962	0.19685	0.0102	0.8968
Factor 13	1.12277	0.18309	0.0087	0.9055
Factor 14	0.93968	0.05378	0.0073	0.9128
Factor 15	0.88591	0.07450	0.0069	0.9196
Factor 16	0.81141	0.11156	0.0063	0.9259
Factor 17	0.69985	0.04486	0.0054	0.9313
Factor 18	0.65499	0.07294	0.0051	0.9364
Factor 19	0.58205	0.03027	0.0045	0.9409
Factor 20	0.55178	0.03941	0.0043	0.9452

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024

Appendix Table A2. Average variance in tasks for top 5 O*NET descriptors for Factor 1 by approach

*O*NET task/skill descriptor*	Guiding, Directing, and Motivating Subordinates	Coordinating the Work and Activities of Others	Developing and Building Teams	Staffing Organizational Units	Coaching and Developing Others
*O*NET descriptor id*	4A4b4	4A4b1	4A4b2	4A4c2	4A4b5
Factor loading	0.9931	0.9362	0.9105	0.8989	0.8440
Latent skill factor	1	1	1	1	1
Factor Interpretation	Non-routine	Non-routine	Non-routine	Non-routine	Non-routine
SOC-based approaches					
SOC 2010	0.015955	0.012791	0.010674	0.014429	0.012386
Task-based approaches					
Structured factor-analysis					
Method 1-109	0.014062	0.011235	0.008948	0.014161	0.010503
Method 2-144	0.014863	0.011550	0.009518	0.015326	0.011039
Unstructured factor-analysis					
Correlated-109	0.009000	0.007920	0.006916	0.011311	0.006914
Correlated-144	0.009774	0.008027	0.007077	0.011704	0.007875

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024.

Appendix Table A3. Average variance in tasks for top 5 O*NET descriptors for Factor 2 by approach

*O*NET task/skill descriptor*	Technology Design	Information Ordering	Inductive Reasoning	Flexibility of Closure	Category Flexibility
*O*NET descriptor id*	2B3b	1A1b6	1A1b5	1A1e2	1A1b7
Factor loading	0.6752	0.6725	0.6575	0.6506	0.6465
Latent skill factor	2	2	2	2	2
Factor Interpretation	Analytical/Cognitive	Analytical/Cognitive	Analytical/Cognitive	Analytical/Cognitive	Analytical/Cognitive
SOC-based approaches					
SOC 2010	0.005887	0.002112	0.003386	0.004278	0.002397
Task-based approaches					
Structured factor-analysis					
Method 1-109	0.007018	0.001683	0.002810	0.003480	0.002169
Method 2-144	0.008138	0.001836	0.003160	0.004163	0.002307
Unstructured factor-analysis					
Correlated-109	0.006570	0.001663	0.002973	0.003681	0.002102
Correlated-144	0.005670	0.001538	0.002507	0.003268	0.001906

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024.

Appendix Table A4. Average variance in tasks for top O*NET descriptors for Factor 3 by approach

*O*NET task/skill descriptor*	Response Orientation	Multilimb Coordination	Performing General Physical Activities	Reaction Time	Operation and Control
*O*NET descriptor id*	1A2b3	1A2b2	4A3a1	1A2c1	2B3h
Factor loading	0.9023	0.8986	0.8713	0.8696	0.8677
Latent skill factor	3	3	3	3	3
Factor Interpretation	Manual	Manual	Manual	Manual	Manual
SOC-based approaches					
SOC 2010	0.010859	0.010853	0.012215	0.012008	0.019450
Task-based approaches					
Structured factor-analysis					
Method 1-109	0.008332	0.009162	0.009240	0.0073430	0.018815
Method 2-144	0.009785	0.009207	0.012353	0.012806	0.015573
Unstructured factor-analysis					
Correlated-109	0.008039	0.009585	0.009483	0.007684	0.018950
Correlated-144	0.008215	0.008559	0.010334	0.009759	0.015962

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024.

Appendix Table A5. Average variance in tasks for top O*NET descriptors for Factor 4 by approach

*O*NET task/skill descriptor*	Self Control	Concern for Others	Deal With Unpleasant or Angry People	Social Orientation	Assisting and Caring for Others
*O*NET descriptor id*	1C4a	1C3b	4C1d2	1C3c	4A4a5
Factor loading	0.8252	0.8197	0.7461	0.7358	0.7136
Latent skill factor	4	4	4	4	4
Factor Interpretation	Interpersonal	Interpersonal	Interpersonal	Interpersonal	Interpersonal
SOC-based approaches					
SOC 2010	0.007040	0.007907	0.015061	0.010682	0.010841
Task-based approaches					
Structured factor-analysis					
Method 1-109	0.008726	0.011999	0.020544	0.014079	0.019175
Method 2-144	0.009041	0.012601	0.020771	0.014929	0.020121
Unstructured factor-analysis					
Correlated-109	0.005049	0.007400	0.012454	0.009806	0.013361
Correlated-144	0.005592	0.008116	0.013053	0.010656	0.014826

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024.

Appendix Table A6. Average variance in tasks for top O*NET descriptors for Factor 5 by approach

*O*NET task/skill descriptor*	Importance of Repeating Same Tasks	Degree of Automation	Consequence of Error	Processing Information	Evaluating Information to Determine Compliance with Standards
*O*NET descriptor id*	4C3b7	4C3b2	4C3a1	4A2a2	4A2a3
Factor loading	0.6742	0.6724	0.4785	0.4757	0.4683
Latent skill factor	5	5	5	5	5
Factor Interpretation	Routine	Routine	Routine	Routine	Routine
SOC-based approaches					
SOC 2010	0.017932	0.013429	0.020106	0.011229	0.012279
Task-based approaches					
Structured factor-analysis					
Method 1-109	n/a	n/a	n/a	0.008641	n/a
Method 2-144	0.021820	0.017354	0.021073	0.010033	0.010668
Unstructured factor-analysis					
Correlated-109	n/a	n/a	n/a	0.009783	n/a
Correlated-144	0.015749	0.014078	0.017940	0.008061	0.008406

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024
n/a indicates this O*NET descriptor was not included in the subset containing 109 variables.

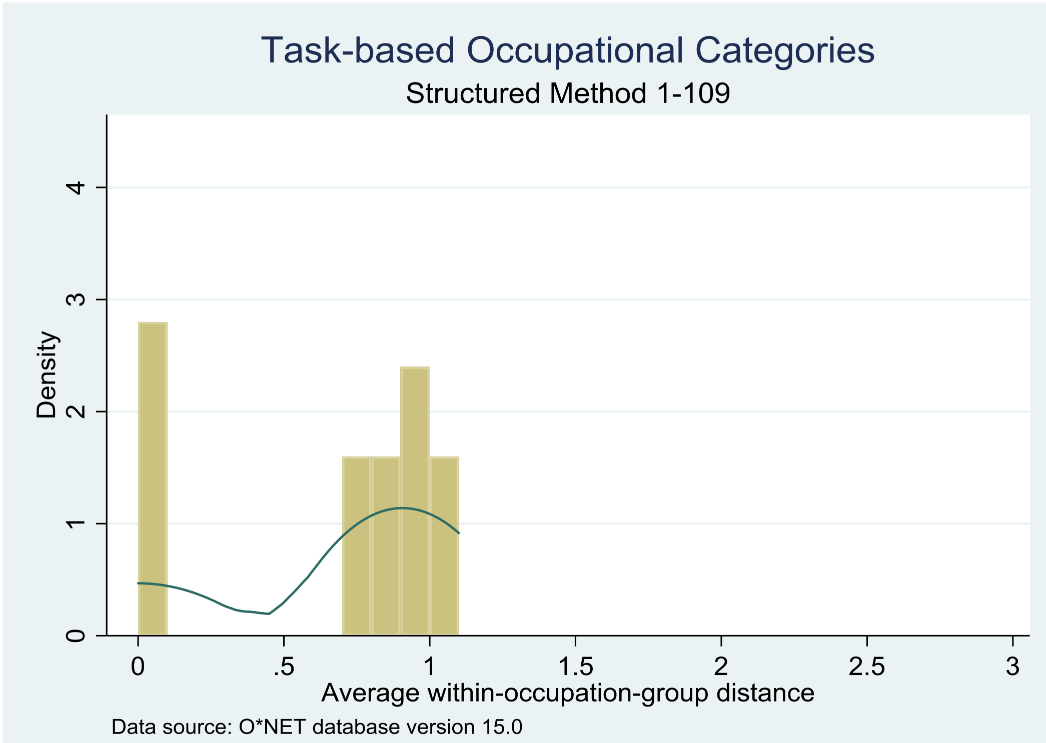
Appendix Table A7. Grand average occupation distance in factor scores (Measure B)

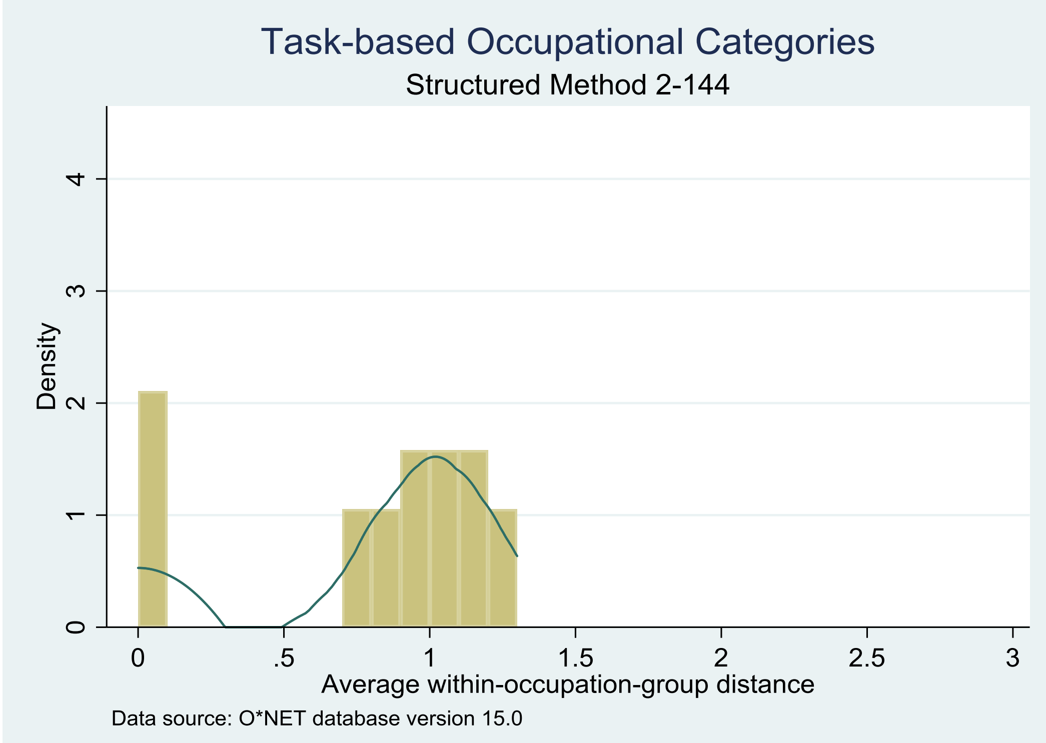
Top panel: 144 O*NET descriptors			
Detailed occupations are assigned the predicted factor score derived from:	Occupational categories are constructed using:		
	SOC-based approach	Task-based approach	
		Structured factor analysis	Unstructured factor analysis
Structured factor-analysis Method 2-144	1.128707	0.916936	0.860402
Unstructured factor-analysis Correlated-144	1.363158	1.494111	1.094561
Bottom panel: 109 O*NET descriptors			
Detailed occupations are assigned the predicted factor score derived from:	Occupational categories are constructed using:		
	SOC-based approach	Task-based approach	
		Structured factor analysis	Unstructured factor analysis
Structured factor-analysis Method 1-109	1.189384	0.929362	0.955579
Unstructured factor-analysis Correlated-109	1.372516	1.413116	1.075908

Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024.

Occupation distance measure B is the Euclidean distance in factors scores between a detailed occupation and the mean of the occupational category. Grand average occupation distance is calculated by (1) calculating occupation distance measure B for each detailed occupation, and (2) calculating the mean of the distance in step 1 across all 485 detailed occupations.

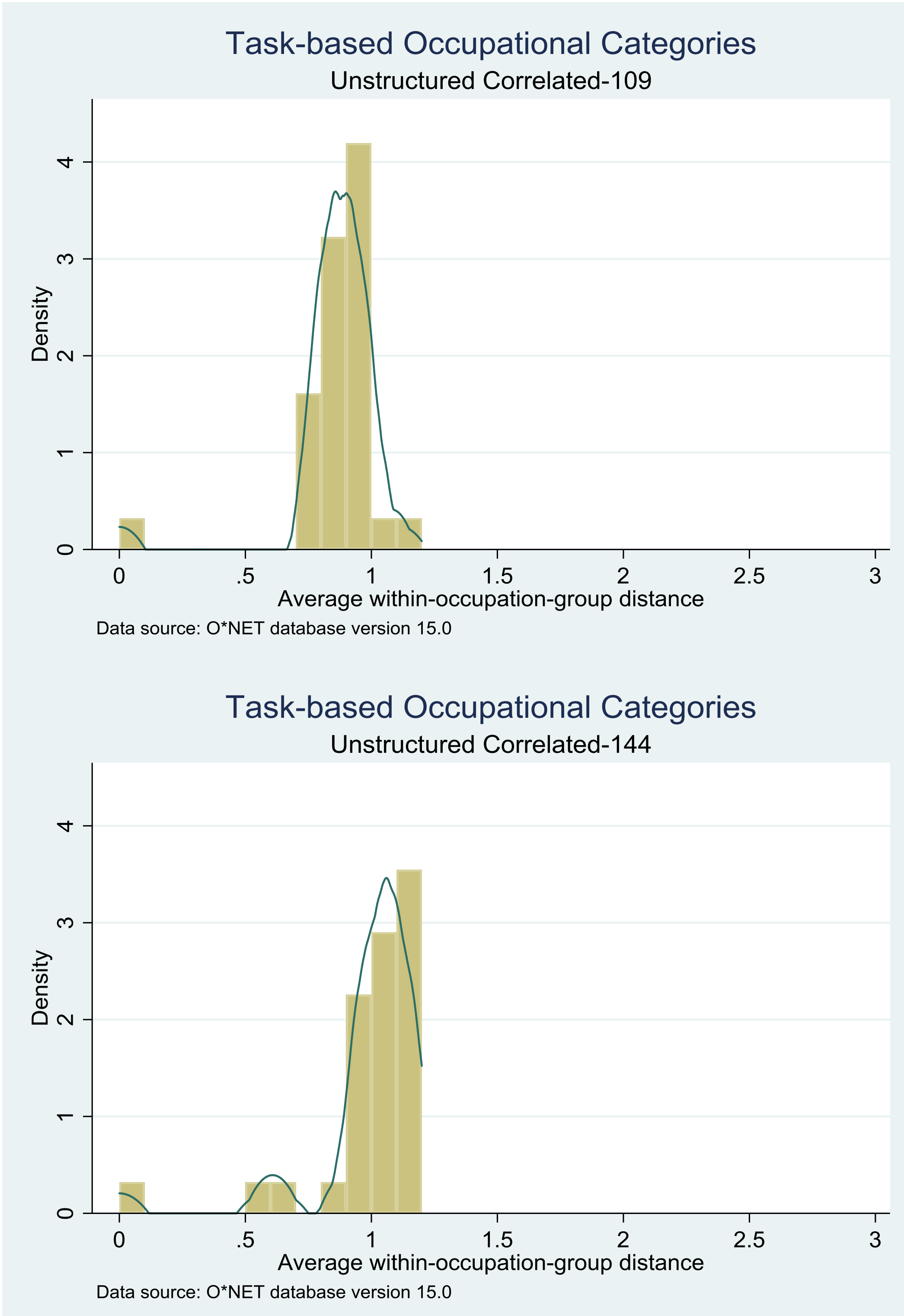
Appendix Figures A1. Histograms of average within-occupation-group distance in tasks, structured task-based approach





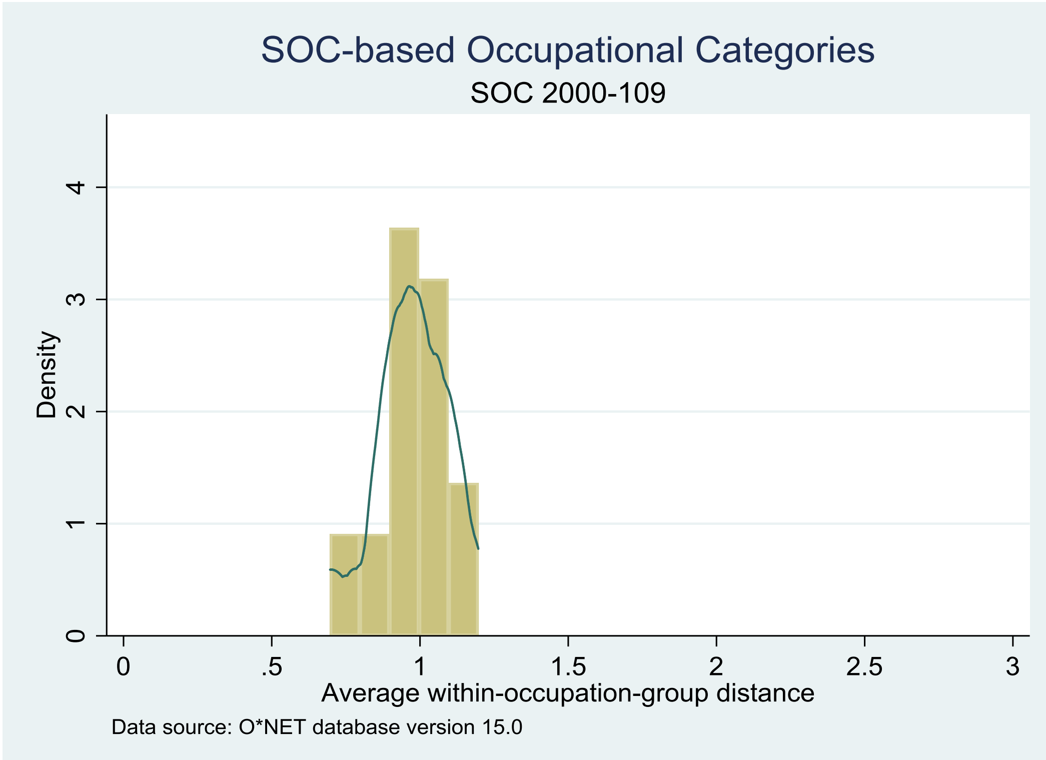
Census DRB release number CBDRB-FY19-CED001-B0024.

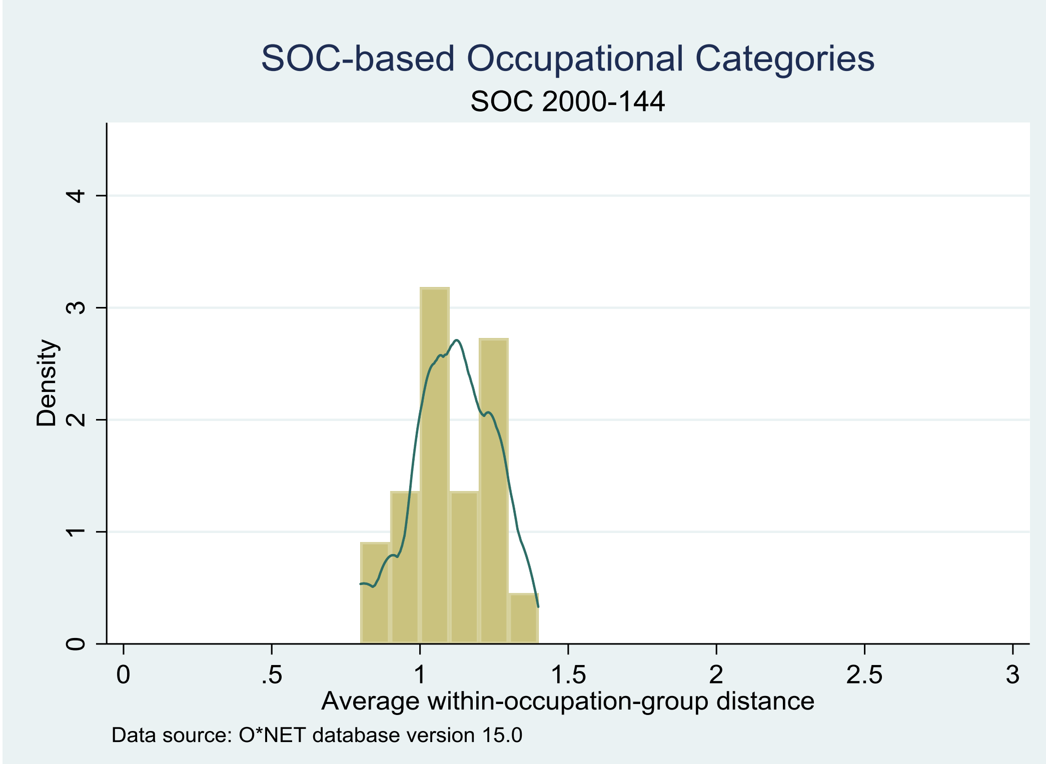
Appendix Figures A2. Histograms of average within-occupation-group distance in tasks, unstructured task-based approach



Census DRB release number CBDRB-FY19-CED001-B0024

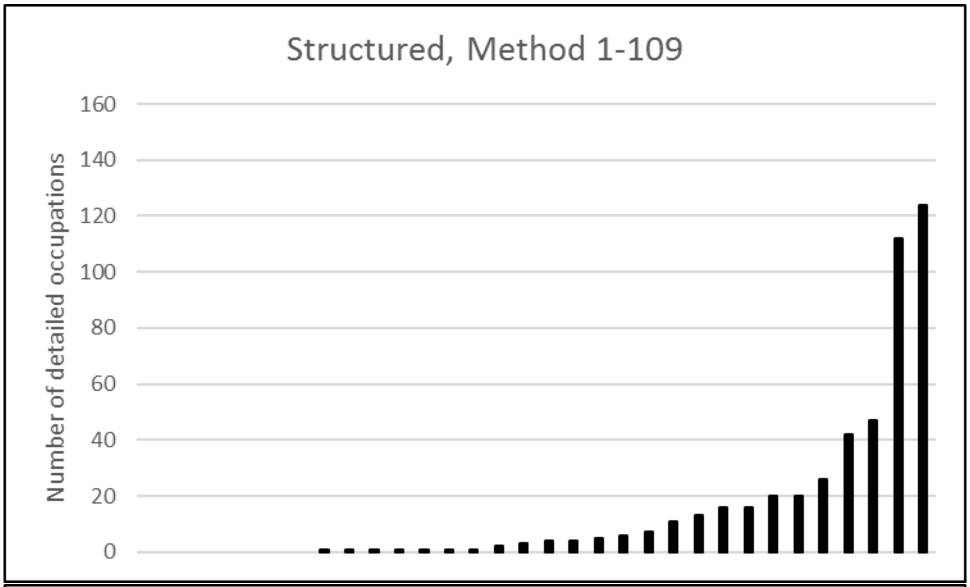
Appendix Figures A3. Histograms of average within-occupation-group distance in tasks, SOC-based approach

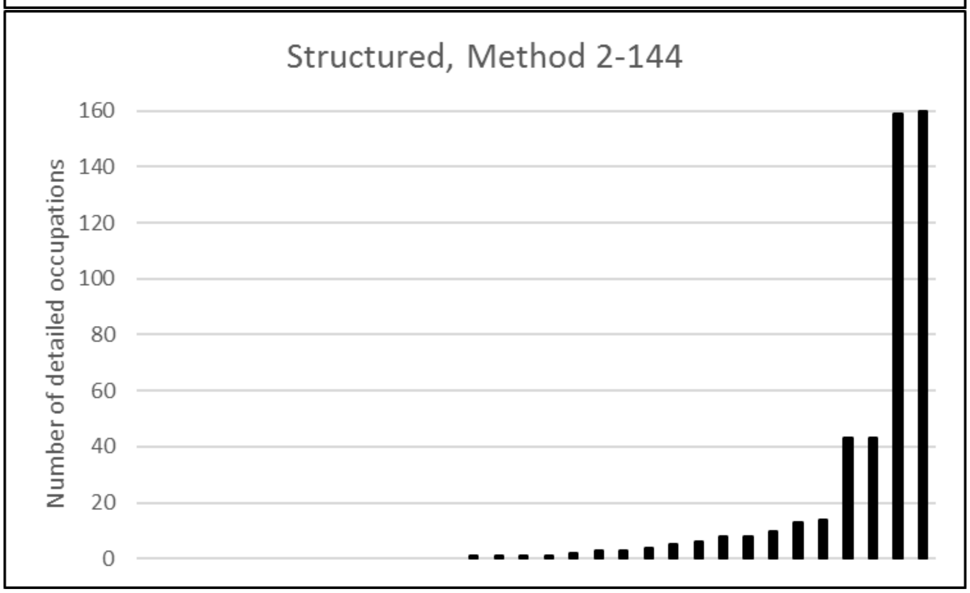




Census DRB release number CBDRB-FY19-CED001-B0024

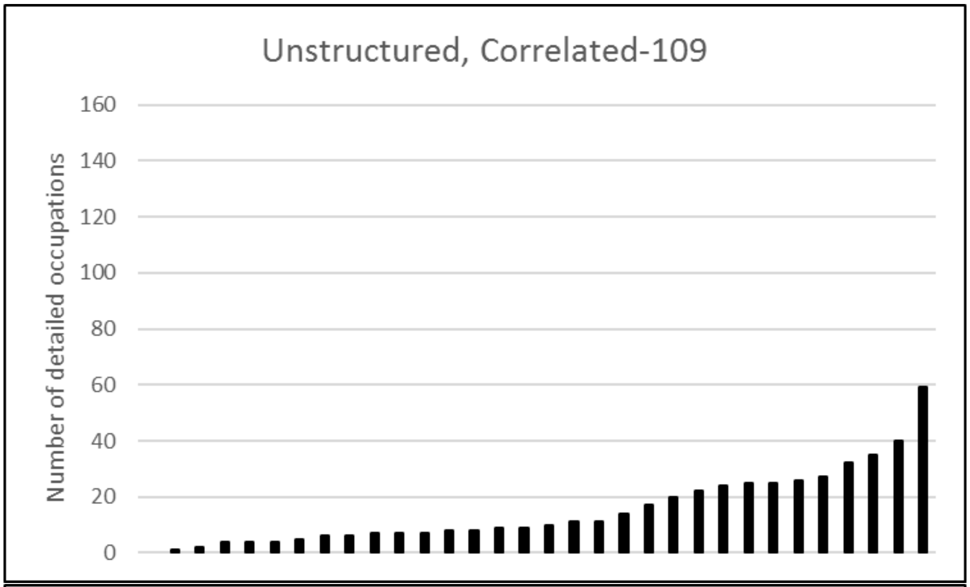
Appendix Figures A4. Ordered bar graphs of the number of detailed occupations in an occupational category, structured task-based approach

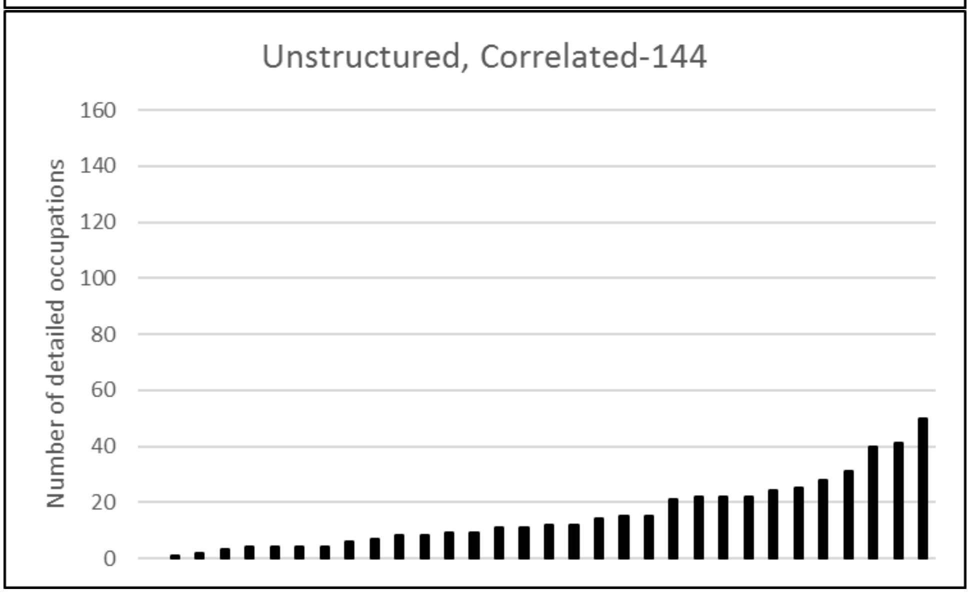




Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024

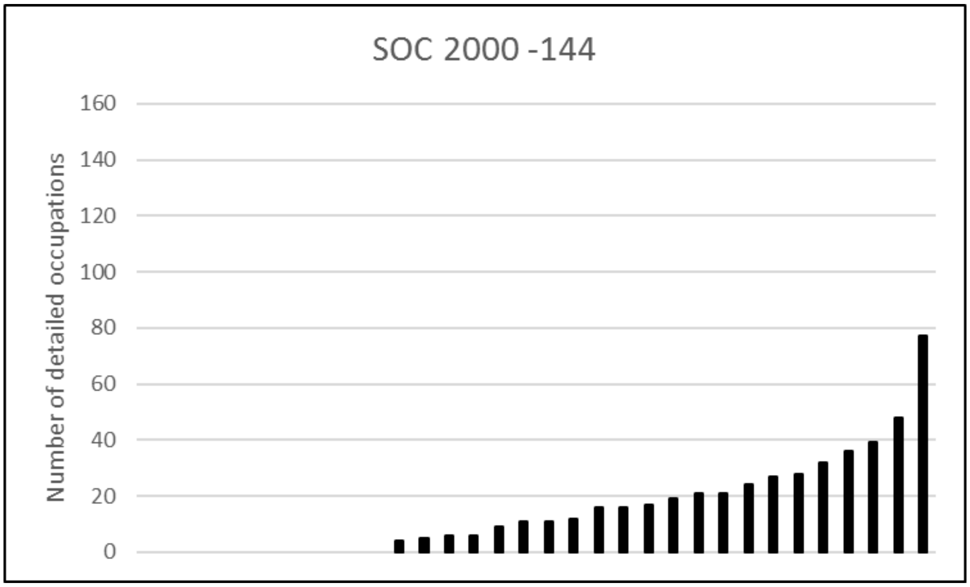
Appendix Figures A5. Ordered bar graphs of the number of detailed occupations in an occupational category, unstructured task-based approach





Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024

Appendix Figures A6. Ordered bar graphs of the number of detailed occupations in an occupational category, SOC-based approach



Data source: O*NET database version 15.0. Census DRB release number CBDRB-FY19-CED001-B0024

Appendix B: Data Appendix for O*NET data files

In order to study the implications of using a task-based approach to quantitatively constructing occupational categories, we needed to create O*NET datasets at both the six-digit 2000 SOC-level and the 2002 Census-level. This appendix provides more details supporting the discussion in section 3.1 of the main text regarding our decision-making in transforming the raw data files from the O*NET version 15.0 database into the SOC-level and Census-level O*NET datasets used in our analysis and applications.

We downloaded O*NET database version 15.0, which was released in July 2010, from the O*NET website on July 19, 2018 (https://www.onetcenter.org/db_releases.html). Our study utilizes five specific data files: Abilities, Workstyles, Skills, Work Activities, and Work Context. O*NET provides documentation that explains that data and occupational information is collected at the O*NET-SOC occupation level. If the O*NET-SOC occupation is directly adopted from the SOC, it is coded at the six-digit SOC-level along with a .00 extension. If the O*NET-SOC occupation is more detailed than the SOC, it is coded at the six-digit SOC-level along with a two-digit extension starting with .01, .02, .03, and so on. In the O*NET data used in our analysis, we have 763 detailed O*NET-SOC level. There is one detailed occupation Legislators (11-1031.00) that is missing all O*NET descriptor rating values and one detailed occupation Mathematical Technicians (15-2091.00) that has many missing O*NET descriptor rating values, so we drop these two detailed occupations from our O*NET dataset. We rescale the descriptor rating values to the interval [0, 1] for all 761 O*NET-SOC occupations using the rescaling formula provided by O*NET.[21]

Next, we create an O*NET dataset at the six-digit SOC level (SOC vintage 2000). We do so in the following way. For cases where there is O*NET descriptor rating values for both the detailed O*NET-SOC level and the six-digit SOC level, we keep the six-digit SOC level descriptor values and drop the more detailed O*NET-SOC level descriptor values. For example, O*NET provide descriptor ratings values for the occupation Medical and Health Services Managers (11-9111.00) and a more detailed occupation Clinical Nurse Specialists (11-9111.01). We keep descriptor values for the former and drop those for the latter since Medical and Health Services Managers (11-9111.00) is a six-digit SOC-level occupation. For cases where there is O*NET descriptor rating values for the detailed O*NET-SOC level and none for the six-digit SOC level, we impute the six-digit SOC level descriptor rating values by taking the mean of the rating values for all of the corresponding detailed occupations at the O*NET-SOC level. For example, there are no O*NET descriptor rating values for the six-digit SOC level occupation Nuclear Technicians (19-4051.00). Yet there are O*NET rating values for two more detailed O*NET-SOC level occupations: Nuclear Equipment Operation Technicians (19-4051.01) and Nuclear Monitoring Technicians (19-4051.02). The means of the descriptor values for these two O*NET-SOC level occupations are the imputed descriptor rating values for the six-digit SOC level occupation, Nuclear Technicians (11-4051.00). For cases where there are no O*NET descriptor rating values for either the detailed O*NET-SOC level or the six-digit SOC level, we drop the six-

[21] Possible original range of values for the different ratings scales are as follows: Level on [0, 7]; Importance on [1, 5]; and Context on [1, 5]. The rescaling formula uses the original rating value, and the lowest and highest possible rating values where the rescaled value = (original-lowest) / (highest-lowest).

digit SOC level occupation from our O*NET dataset. For example, there are no O*NET descriptor ratings values for the detailed occupation Legislators (15-2091.00) so this occupation is dropped from our O*NET dataset. This occurs mostly for detailed SOC occupations with titles containing words like “miscellaneous”, “all other”, or “not elsewhere classified” (n.e.c.). For example, there are no O*NET descriptor ratings values for the detailed occupation Production workers, all other (51-9999.00) so this occupation is dropped from our O*NET dataset. This first collapsing step results in a balanced O*NET descriptor dataset for 757 detailed occupations at the six-digit SOC-level.

While Census Occupation Codes are based on the SOC, sometimes detailed occupations are collapsed into broad occupations due to collectability issues. So we also create an O*NET dataset at the 2002 Census Occupation Code level (henceforth, COC2002). For cases when more than one detailed occupation at the 2000 SOC level is paired with one detailed/broad occupation at the COC2002 level, we impute the COC2002 level descriptor rating values by taking the mean of the rating values for all of the corresponding detailed occupations at the SOC level. For example, in the COC2002 list 2300 is equivalent to the SOC broad occupation, Preschool and Kindergarten Teachers (SOC 25-2010), which collapses two detailed occupations Preschool Teachers, except Special Education (SOC 25-2011) and Kindergarten Teachers, Except Special Education (SOC 25-2012). In our O*NET dataset at the COC2002 level, the O*NET descriptor values for the two detailed occupations (SOC 25-2011 and 25-2012) are averaged to obtain the mean O*NET descriptor rating values for Preschool and Kindergarten Teachers (SOC 25-2010). This second collapsing step results in a balanced O*NET descriptor dataset for 490 detailed occupations at the COC2002 level.

Appendix C: Data Appendix for Replication

Following the PSID analysis in Blau and Kahn (2017), we restrict the sample to include individuals aged 25-64 at the time of their 2008 SIPP survey who have worked at least 26 weeks and have a wage value greater than or equal to $2/hour. Observations with missing data for variables of interest are dropped. We do not use weights in estimating our full specification regression models. Education variables include years of schooling completed, an indicator for having a bachelor's degree, and an indicator for having a graduate degree. Experience variables include the number of years with positive earnings in the IRS/SSA Summary Earnings Record (SER) and its square. There are also indicators for each Census region, for race and Hispanic status, and for unionization. We construct industry dummy variables as it is done in the analysis in Blau and Kahn (2017).[22]

Upon a closer examination of the details in the online data appendix for Blau and Kahn (2017), we noticed that the occupational categories they used were predominantly two-digit SOC-based groupings with a few adjustments. For example, the occupational category, Post-secondary Educators, consisted of a single detailed occupation, Post-secondary Teachers (25-1000), instead of being grouped with other detailed occupations in the two-digit SOC group (25-0000), Education, Training, and Library Occupations. In another example, detailed occupations such as Lawyers (23-1011); Judges, magistrates, and other judicial workers (23-1020); Physicians and surgeons (29-1060); and Dentists (29-1020) are pulled out of their respective two-digit SOC group and then combined to form the occupational category, Lawyers, Judges,

[22] We used the materials found in the online data appendix for Blau and Kahn (2017) https://www.aeaweb.org/articles?id=10.1257/jel.20160995.

Physicians and Dentists. Some of the adjustments may reduce the within-group variance in earnings. So we use two sets of SOC-based occupational categories: the true two-digit SOC groups and the groups in Blau and Kahn (2017).