Behavior Research Methods (2024) 56:417–432
https://doi.org/10.3758/s13428-022-02044-7

Visualization of latent components assessed in O*Net occupations (VOLCANO): A robust method for standardized conversion of occupational labels to ratio scale format

Ju-Chi Yu[1,2] · H. Moriah Sokolowski[4] · Kirthana S. Rao[2] · Luke E. Moraglia[2] · Soudeh A. Khoubrouy[2,3] · Hervé Abdi[2] · Brian Levine[4,5,6]

Accepted: 30 November 2022 / Published online: 25 January 2023

Abstract
Occupations are typically characterized in nominal form, a format that limits options for hypothesis testing and data analysis. We drew upon ratings of knowledge, skills, and abilities for 966 occupations listed in the US Department of Labor's Occupational Classification Network (O*NET) database to create an accessible, standardized multidimensional space in which occupations can be quantitatively localized and compared. Principal component analysis revealed that the occupation space comprises three main dimensions that correspond to (1) the required amount of education and training, (2) the degree to which an occupation falls within a science, technology, engineering, and mathematics (STEM) discipline versus social sciences and humanities, and (3) whether occupations are more mathematically or health related. Additional occupational spaces reflecting cognitive versus labor-oriented categories were created for finer-grained characterization of dimensions within occupational sets defined by higher or lower required educational preparation. Data-driven groupings of related occupations were obtained with hierarchical cluster analysis (HCA). Proof-of-principle was demonstrated with a real-world dataset (470 participants from the Nathan Kline Institute – Rockland Sample; NKI-RS), whereby verbal and non-verbal abilities—as assessed by standardized testing—were related to the STEM versus social sciences and humanities dimension. Visualization of Latent Components Assessed in O*Net Occupations (VOLCANO) is provided to the research community as a freely accessible tool, along with a Shiny app for users to extract quantitative scores along the relevant dimensions. VOLCANO brings much-needed standardization to unwieldy occupational data. Moreover, it can be used to create new occupational spaces customized to specific research domains.

Keywords Occupation · Vocation · Principal component analysis · Cognitive ability · Multivariate analyses

Introduction

Understanding an occupation (i.e., a person's regular work or profession) is relevant to research questions across many disciplines, including psychology, neuroscience, economics, political science, education, sociology, health care, and ageing. These questions often require retrospective analyses relating occupational practice to other measures, such as—among others—health status, cognitive abilities, and sex or gender differences. By contrast, vocational psychologists use

Ju-Chi Yu and H. Moriah Sokolowski contributed equally.

✉ Ju-Chi Yu
Ju-Chi.Yu@camh.ca

✉ H. Moriah Sokolowski
h.moriah.sokolowski@gmail.com

✉ Hervé Abdi
herve@utdallas.edu

✉ Brian Levine
blevine@research.baycrest.org

[1] Campbell Family Mental Health, Centre for Addiction and Mental Health, Toronto, Canada

[2] School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

[3] Department of Bioengineering, The University of Texas at Dallas, Richardson, TX, USA

[4] Rotman Research Institute, Baycrest Centre, Toronto, Canada

[5] Department of Psychology, University of Toronto, Toronto, Canada

[6] Department of Medicine (Neurology), University of Toronto, Toronto, Canada

prospective analysis of personality traits, abilities, and interests to predict placement or select candidate (e.g., Hartman & Betz, 2007; Holland, 1997; Larson et al., 2002; Ralston et al., 2004)—a procedure that has been criticized due to low empirical support (Fouad & Kozlowski, 2019; Savickas, 2001).

Tools that can convert occupations into quantified dimensions of mental traits and skills—as well as contextual factors (e.g., sex/gender and economic access)—are required to test hypotheses regarding characteristics of individuals whose occupation is known. Researchers across disciplines have strived to develop an "occupational space" that can be linked to various research questions. For instance, cognitive neuropsychologists have identified key occupational attributes that relate to the location of atrophy in frontotemporal lobar degeneration in order to understand how long-term engagement in occupations associates with individual differences in the emergence of neurodegeneration (e.g., Spreng et al., 2010). Meanwhile, cognitive neuroscientists have assessed how occupations associate with structural brain health (e.g., Habeck et al., 2019) and gerontologists have identified key occupational factors in order to identify specific physical traits of occupations that could predict healthy ageing (e.g., Burzynska et al., 2019). While there is some overlap in the cognitive factors identified by these different research teams (e.g., a factor relating to the occupational complexity), these methods used to identify occupational factors differ across teams and disciplines, a problem making related findings challenging to compare.

There are several methodological barriers to developing continuous measures of occupation characteristics. Occupation is a categorical variable with a large number of titles (i.e., modalities), including multiple titles describing the same occupation (e.g., medical doctor vs. physician). Because occupations can be described with varying degrees of specificity (e.g., physician vs. cardiologist), each occupational title exists within a complex hierarchical structure. Additionally, each occupation requires a different mix of knowledge, qualifications, and abilities, and this makes it challenging to categorize occupations into subgroups.

The dictionary of occupational titles (DOT)—first created in 1930 by the United States Department of Labor (DOT: United States Department of Labor, 2006)—is the earliest and best-known occupation taxonomic system. In 1995, the DOT was replaced by the Occupational Classification Network (O*NET) database (US Department of Labor, 2019a), reducing the DOT's 28,800 occupations to 966 occupations grouped using a conceptual framework called the Content Model (see Peterson et al., 2001). O*NET's sampling methodology and occupational assignment improved upon the DOT, enhancing application in social sciences (Handel, 2016) (Peterson et al., 2001). The O*NET Content Model provides a framework that identifies the most important types of information about occupations and integrates this information into a theoretical, empirically validated, model. The O*NET content model includes both worker-oriented as well as job-oriented traits. The O*NET content model describes and categorizes these distinguishing characteristics of occupations and provides a set of standardized, measurable variables that represent key features of occupations. Worker-oriented variables, including the specific knowledge, skills, and abilities associated with particular occupations, have long been used to explore how particular traits relate to different aspects of occupation at the individual level (e.g., Burrus et al., 2013).

Both DOT and O*NET data have also been subjected to dimension reduction techniques in order to extract components accounting for fundamental occupational traits, such as occupational complexity, people versus things, and physical demands (Clark, 2002; Hadden et al., 2004; Hanson et al., 1999; Levine, 2003; Shu et al., 1996). Similarly, clustering approaches have been applied to reduce the complexity of O*NET data (e.g., Nolan et al., 2011; Slaper, 2014). Although several studies have reduced O*NET data into accessible, sharable, and actionable data (Indiana Department of Workforce Development Research and Analysis Division & Indiana Business Research Center, 2011), these analyses have been tailored to select samples or occupational categories and are not flexible or modifiable to suit different research questions. A standardized, accessible, and flexible system for quantification of occupation characteristics is required to scale the application of the rich O*NET database across research domains.

We created a method for the derivation of quantified dimensional scores characterizing O*NET occupations for use by the research community. We applied principal component analysis (PCA; Abdi & Williams, 2010; Eckart & Young, 1936; Hotelling, 1933) to the ratings of knowledge, skills, and abilities associated with each occupation from O*NET to identify dimensions (also called components) that capture patterns in the data. We then used these components as the input to hierarchical cluster analyses (HCA; Bridges, 1966) to identify categorical clusters of occupations and their associated traits as indicated by the ratings. Visualizations of the occupational space were labeled by these clusters as well as by the overall required level of preparation (i.e., 'Job Zone', which quantifies the required level of education, related experience, and on-the-job training, see https://www.onetonline.org/help/online/zones). Follow-up PCAs with HCAs within different Job Zones were conducted to identify finer-grained occupational components unaccounted for by education and socioeconomic status. We also provide an illustrative example where we use Visualization of Latent Components Assessed in O*Net Occupations (VOLCANO) to understand the link between cognition and occupation in a sample of healthy adults. All analyses, data, and clustering are available through the Open Science Framework (OSF), GitHub (https://github.com/juchiyu/OccupationPCAs), and an original Shiny app (i.e., an R-based interactive web app; the codes and link are included in the same GitHub repository).

Table 1 Sample data of occupations and associated traits

Occupations	Traits					
	Abilities		Knowledge		Skills	
	Arm Hand Steadiness	Deductive Reasoning	Chemistry	Production & Processing	Active Listening	Programming
Job Zones 1–3						
Bartenders	3.00	3.00	1.87	1.69	3.12	0.25
Electricians	3.00	3.88	1.50	2.39	3.00	0.00
Maids and Housekeeping Cleaners	2.12	2.38	0.97	1.12	2.25	0.00
Sales Agent, Financial Services	1.00	4.25	0.07	0.99	4.00	0.75
Job Zones 4 – 5						
Chemists	3.12	4.50	6.10	3.85	3.75	1.75
Computer Programmers	1.12	3.88	0.29	2.56	4.00	4.88
Fundraisers	0.00	4.25	0.06	0.56	3.88	1.00
Lawyers	0.38	4.38	0.38	1.46	4.75	0.75

Occupational traits were selected from 120 variables across 966 occupations for the purposes of illustration. Trait ratings range from "not important" (0) to "extremely important" (5)

Method

Data collection and extraction

The data included in this study were obtained from a public database available from the United States Department of Labor Standard Occupational Classification Network (O*NET: US Department of Labor, 2019a). O*NET is a Standard Occupational Classification based system that organizes work into 966 occupations and associated traits as determined from surveys of workers (US Department of Labor, 2019c).

In our analyses, each occupation is treated as an "observation" described by scores on different measures. For each occupation within the O*NET database, we extracted 120 trait variables (e.g., trunk strength ability, physics knowledge, and clerical skills) that were described with Likert-scaled ratings ranging from "not important" (0) to "extremely important" (5) concerning abilities (52 variables), knowledge (33 variables), and skills (35 variables). The current study did not include job-oriented traits (US Department of Labor, 2019b), because the key aim of this study is to uncover the association between individual abilities (e.g., cognitive, behavioral, neural) and worker-level occupational attributes. These data (966 occupations described by 120 traits) were used as input for all subsequent analyses (see Table 1 for examples).

Analyses

A *general occupation PCA* was applied to the 966 occupations described by 120 rated traits. The first component of this general occupation PCA was dominated by Job Zone level (i.e., a five-level grouping variable that reflects the preparation required; specifically, the amount of education, related experience, and on-the-job training needed to do the work). To identify components that explain variance over and above the effect of Job Zone effect, we conducted follow-up PCAs to identify components only within Job Zones 4 and 5 (391 occupations requiring extensive preparation; the *Job Zones 4–5 PCA*) and Job Zones 1–3 (575 occupations requiring relatively less preparation; the *Job Zones 1–3 PCA*). Data were analyzed using R, version 4.1.1 (R Core Team , 2021), with packages *ExPosition*, version 2.8.23 (Beaton et al., 2014), *stats*, version 4.1.1 (R Core Team, 2021), *ggplot2*, version 3.3.5 (Wickham, 2016), *ggdendro*, version 0.1.22 (de Vries & Ripley, 2020), *dendextend*, version 1.15.1 (Galili, 2015), and *tidyverse*, version 1.3.1 (Wickham et al., 2019).

Principal component analysis (PCA)

We performed PCA on the preprocessed data where each trait was mean-centered (i.e., the mean of each trait is now equal to 0) across occupations (with the *ExPosition* R package, Beaton et al., 2014; scaling—a.k.a., normalization, such as using *Z*-scores—was not used because all traits were rated on the same scale)[1]. PCA creates a set of orthogonal

[1] We did not perform a rotation here, because rotation is recommended when the variables are normalized (i.e., so that the average eigenvalue is unity) and works best when the pattern of eigenvalues clearly indicates the number of components to keep (see Abdi & Williams, 2010). However, when we performed a Varimax rotation with the number of dimensions equal to the number of significant components—as determined by the permutation test —, we obtained results highly similar to the un-rotated PCA reported here.

Table 2 Descriptions of occupation cluster labels

	Occupation Clusters	
	Label	Description
General	ArtsCom	Arts and Communication Specialists
	BioTech	Biological Technicians
	Construction	Construction Workers and Craftsmen
	EnviroSaftey	Environment and Safety Managers
	EnviroSustain	Environment and Sustainability Managers
	FinMedLglClerks	Financial, Medical and Legal Clerks
	FinMedLglManag	Financial, Medical and Legal Managers
	FinMedLglSpec	Financial, Medical and Legal Specialists
	GovtComManag	Government and Community Managers
	ITSpec	Computer and Informatics Specialists
	Labor	Physical Labor Workers
	MathSpatial	Mathematicians and Spatial Scientists
	MedPrac	Medical Practitioners
	ObjectTech	Object Technicians
	SciEngi	Scientists and Engineers
	Service	Service Workers
	SocSci	Social Scientists and Practitioners
	TechOp	Technicians and Operators
Job Zones 4–5	AltThrpy	Alternative Therapy
	BusGovt	Business and Government
	CompSci	Computer and Informatics
	Engi	Engineering
	Enviro	Environment
	MedGen	Medical Science, Generalized
	MedSpec	Medical Science, Specialized
	SalesLog	Sales and Logistics
	SciMaths	Science and Mathematics
	SocSci	Social Science
Job Zones 1–3	Clerk	Office Clerks
	EngEnviroTech	Engineering and Environmental Technicians
	HealthTech	Health Technicians
	LaborFineMotor	Labor Workers, Requiring Fine Motor Skills
	LaborGrossMotor	Labor Workers, Requiring Gross Motor Skills
	Managerial	Managerial Workers
	Production	Production Workers
	PublicSafe	Public Safety Specialists
	ServiceSocial	Service Workers, Social Sector
	ServiceTech	Service Workers, Technical Sector

(i.e., uncorrelated) variables called *principal components* (Abdi & Williams, 2010; Hotelling, 1933; Pearson, 1901). The scores of each principal component are called *component scores* and are obtained as a linear combination of the original variables (i.e., here traits) with coefficients—called *loadings*—that indicate the importance of each original variable in the combination. For each component, the amount of explained variance in the data is measured by the variance of its component scores—called the *eigenvalue* of this component. In PCA, these components are ordered (from the largest to the smallest) according to their eigenvalues.

In PCA, a component is reliable when it explains a significant amount of variance. This significance was evaluated by a permutation test obtained by (1) randomly permuting the data within each variable, (2) computing a PCA on the permuted data table, (3) repeating this process

many (i.e., 1000) times, and (4) generating the probability distribution of each eigenvalue from the PCAs of these permuted datasets. From each distribution, the proportion of the permuted eigenvalues that are larger than the observed eigenvalue gives the probability associated with (i.e., the *p* value of) this eigenvalue (Abdi & Williams, 2010; Buja & Eyuboglu, 1992; Reddon, 1984).

In PCA, these components are interpreted by inspecting, one component at a time, their patterns of component scores and loadings, and this process is facilitated by drawing maps (i.e., scatterplots) that plot the loadings of two components (typically Components 1 and 2) against each other. In these maps, the correlation between two traits is estimated by the angle that these two traits form with the origin of the map: Two traits with a small angle are *positively* correlated, two traits with a right angle are *orthogonal* (i.e., uncorrelated), and two traits with a large obtuse angle are *negatively* correlated. Here, loadings were scaled to have the same variance as the component scores and therefore called "scaled loadings" (by contrast with the usual loadings that have unitary variance).

Finally, to reveal the structure of the occupations in the same component space, we used the component scores as coordinates to create scatterplots of the occupations. In these maps, the similarity between occupations is evaluated by their *distances*: Occupations near each other require similar skills, whereas occupations far from each other require distinct skills (note that this procedure differs from the plots of the traits whose interpretation relies on the *angle* between pairs of traits). We labeled these components according to how both traits and occupations are distributed. The labels were selected for the sake of efficient communication, recognizing that they do not necessarily reflect the nuances of each component.

Hierarchical cluster analysis (HCA)

Hierarchical cluster analyses (HCA; Bridges, 1966)[2] were performed on both the occupations' component scores and the traits' scaled loadings for the three PCAs (i.e., general, Job Zones 1–3, Job Zones 4–5) using the significant components of each PCA. The HCAs were conducted using the R-function `hclust` (R Core Team, 2021), and the clusters were extracted using Ward's minimum variance (Ward, 1963). Homogeneous clusters were defined such that the numbers of occupations or traits were roughly equivalent (Fig. 1; see the detailed list of occupations and traits in Supplementary S1, S3, and S4).

[2] A *K*-mean cluster analysis was also implemented to extract clusters of occupations and occupational traits. However, the resulting clusters were not as homogeneous as the clusters from HCA and thus were not kept for the current study.

Visualizations

Component scores and scaled loadings were grouped and labeled using clusters derived from the corresponding HCAs. The component scores are also grouped and labeled using the Job Zone variable from O*NET to assess the degree to which a given component corresponds to the overall level of education and preparation required. An interactive, publicly available (R-based) Shiny app (available at https://github.com/juchiyu/OccupationPCAs) is provided so that others can re-run analyses and plot the data, adjusting the parameters (e.g., the number of clusters, clustering method, occupation vs trait clustering) according to their needs.

Results

General occupation PCA

The general occupation PCA identified seven significant components, which, together, explained 77.24% of the variance. HCA identified 18 occupation clusters and nine trait clusters (Fig. 1A and Supplementary S1). The occupation spaces of the first three components are shown in Figs. 2 and 3. The first component, explaining 37% of the variance, differentiated labor-intensive occupations (e.g., stonemasons, logging equipment operators) from education-intensive occupations (e.g., industrial-organizational psychologists, political science teachers), with the scaled loadings of traits reflecting physical versus cognitive abilities (Fig. 2B and Supplementary S5, horizontal axes). Accordingly, this component corresponded to the five O*NET Job Zone groups (Fig. 2A; horizontal axis). Because this component was largely determined by the amount of preparation required (including education, related experience, and on-the-job training), we labeled it "Preparation." This component likely reflects what has previously been labeled "occupational complexity" or "substantive complexity" (Crouter et al., 2006; Gadermann et al., 2014; Hadden et al., 2004; Smart et al., 2014).

The second component explained 19% of the variance and differentiated occupations in STEM (e.g., robotic, marine, biomedical engineers) from occupations that are non-STEM (e.g., models, telemarketers, coatroom attendants), with the scaled loadings of traits reflecting engineering, technology, and natural sciences (Fig. 2B and Supplementary S5, vertical axes). These scaled loadings of traits such as fine arts and philosophy on the low end of this component were close to 0 and therefore did not contribute strongly to this component. We therefore simply called this component "STEM."

The third component explained 9% of the variance and differentiated occupations in medicine, health science, and social science (e.g., nurse practitioners,

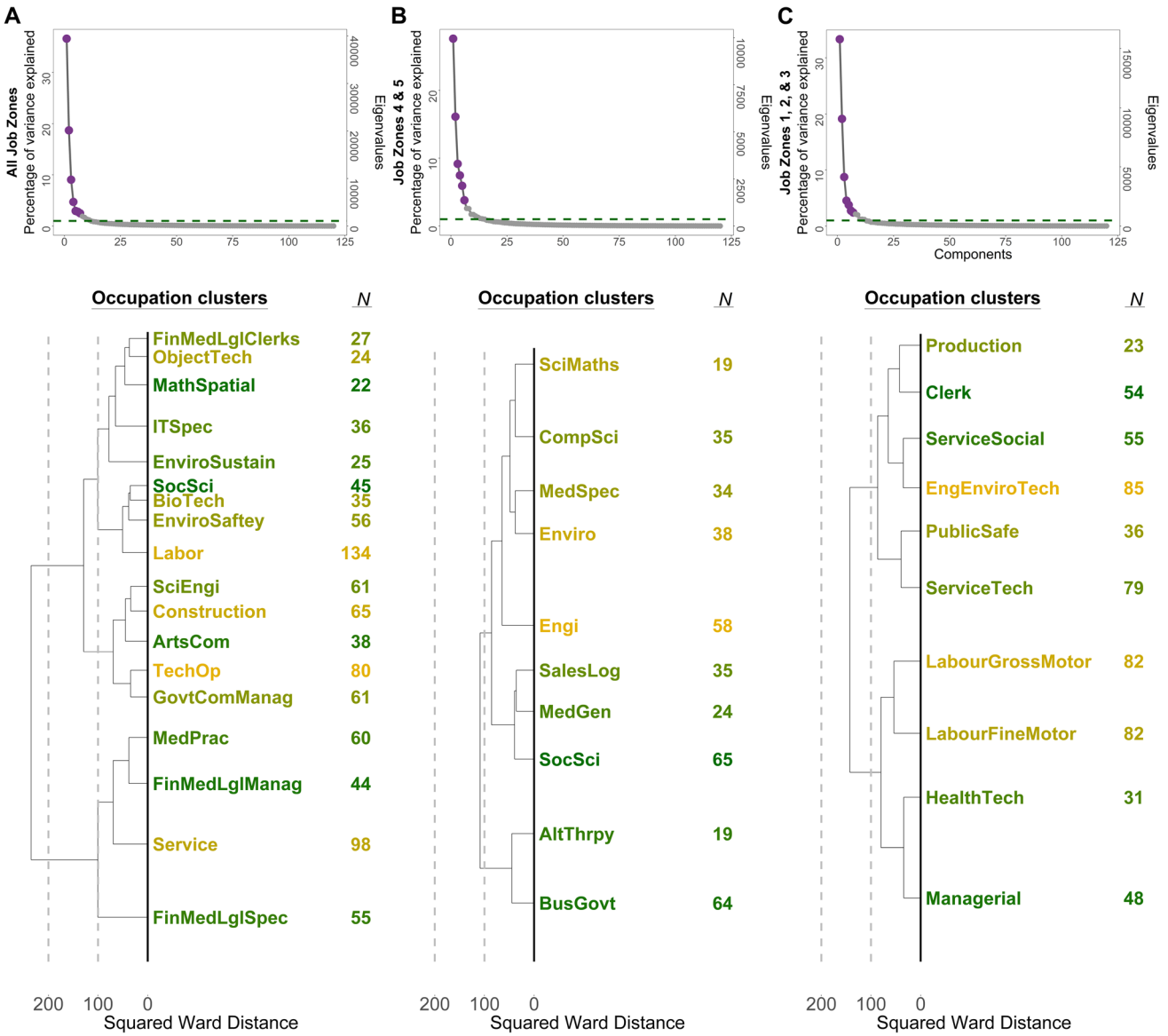

Fig. 1 Descriptive information of the General, the Job Zones 4–5, and the Job Zones 1–3 PCAs. *Note*. Descriptive information for the General occupation (**A**), Job Zones 4–5 (**B**) and Job Zones 1–3 (**C**) PCAs and HCAs. *Top*: Scree plots depicting the percentage of variance explained by each component. The *green dashed lines* indicate the Kaiser criteria, and the *purple dots* indicate significant components as determined by permutation tests. *Bottom*: Tree diagrams depicting the hierarchical structure of the clustering of occupations from the HCA (N = number of occupations). Clusters were colored using a gradient of the component scores / scaled loadings of the first components (i.e., *yellow-green* reflects positive-negative loadings on the first component for occupation; see Figs. 2, 3, 4 and 5 for color gradients). See Supplementary S1, S2, and S4 for the tree diagram depicting the hierarchical structure of the clustering of traits. Details of abbreviations for the occupation clusters are listed in Table 2

clinical nurse specialists, police officers) from computer science and engineering (e.g., hardware engineers, aerospace engineers, mechanical drafters), with the scaled loadings of traits anchored by engineering and technology versus social sciences, humanities, and natural sciences (Fig. 3B and Supplementary S6, vertical axes). We labeled this component "Health versus Computational Science."

Job zones 4–5 PCA

Job Zones 4–5 PCA identified six significant components, which, together, explained 70% of the total variance. HCA was conducted on the first four of these components (which, together, explained 60% of the variance because the last two components did not yield interpretable clusters,

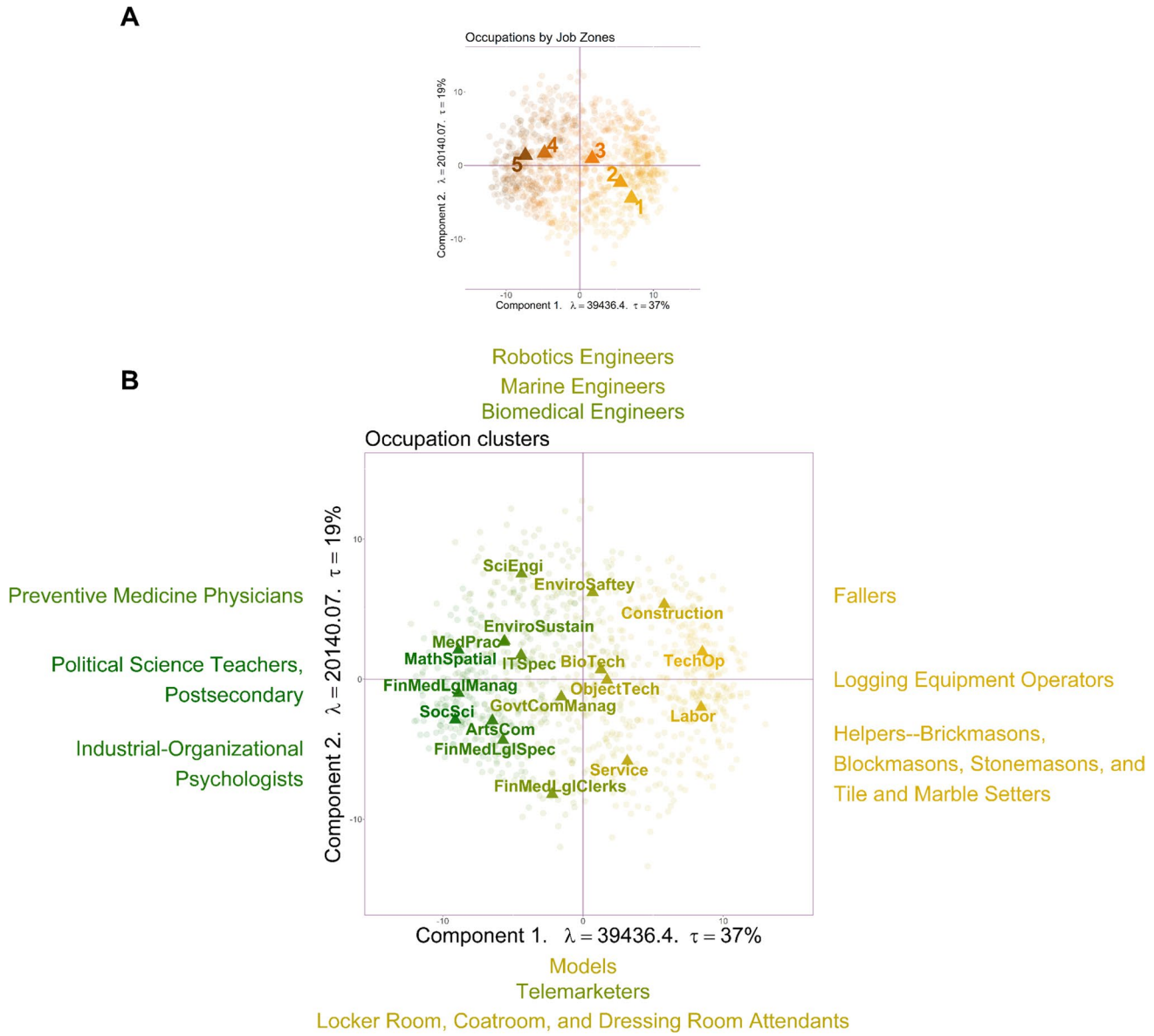

Fig. 2 First two components (Preparation and STEM) from the General occupation PCA. *Note.* Panel **A** illustrates the degree to which each component corresponds to mean Job Zone ratings. Anchors indicate occupations (**B**) with the highest contributions (e.g., [Tree] Fallers and extent flexibility contributed strongly and positively to Preparation [Component 1]). Occupation component scores (**B**) are labeled by clusters derived from the HCA. The horizonal axis in all plots represents Preparation (Component 1), and the vertical axis represents STEM (Component 2)

Supplementary S2). This HCA identified 18 occupation clusters and 11 trait clusters (Fig. 1B and Supplementary S3). Figure 4 shows the occupation space of the first two components of the Job Zones 4–5 PCA. The preparation-level factor observed in general occupation PCA no longer dominated the first component, likely because the Job Zones 4–5 PCA is restricted to occupations requiring considerable to extensive preparation level (see Fig. 4A). Job Zones 4–5 contain occupations that rely on cognitive abilities (e.g., coordinating, supervising, managing, etc.).

The first and second components of Job Zones 4–5 PCA (Fig. 4) resembled the second and third components of the general occupation PCA (Fig. 3). The first component, explaining 27% of the total variance, differentiated STEM occupations (e.g., manufacturing, marine, robotic engineers) from occupations in the humanities, social sciences, and particularly teachers (e.g., English language and literature teachers, history teachers), with the scaled loadings of traits reflecting science and engineering versus social science, humanities, and communication (Fig. 4B and

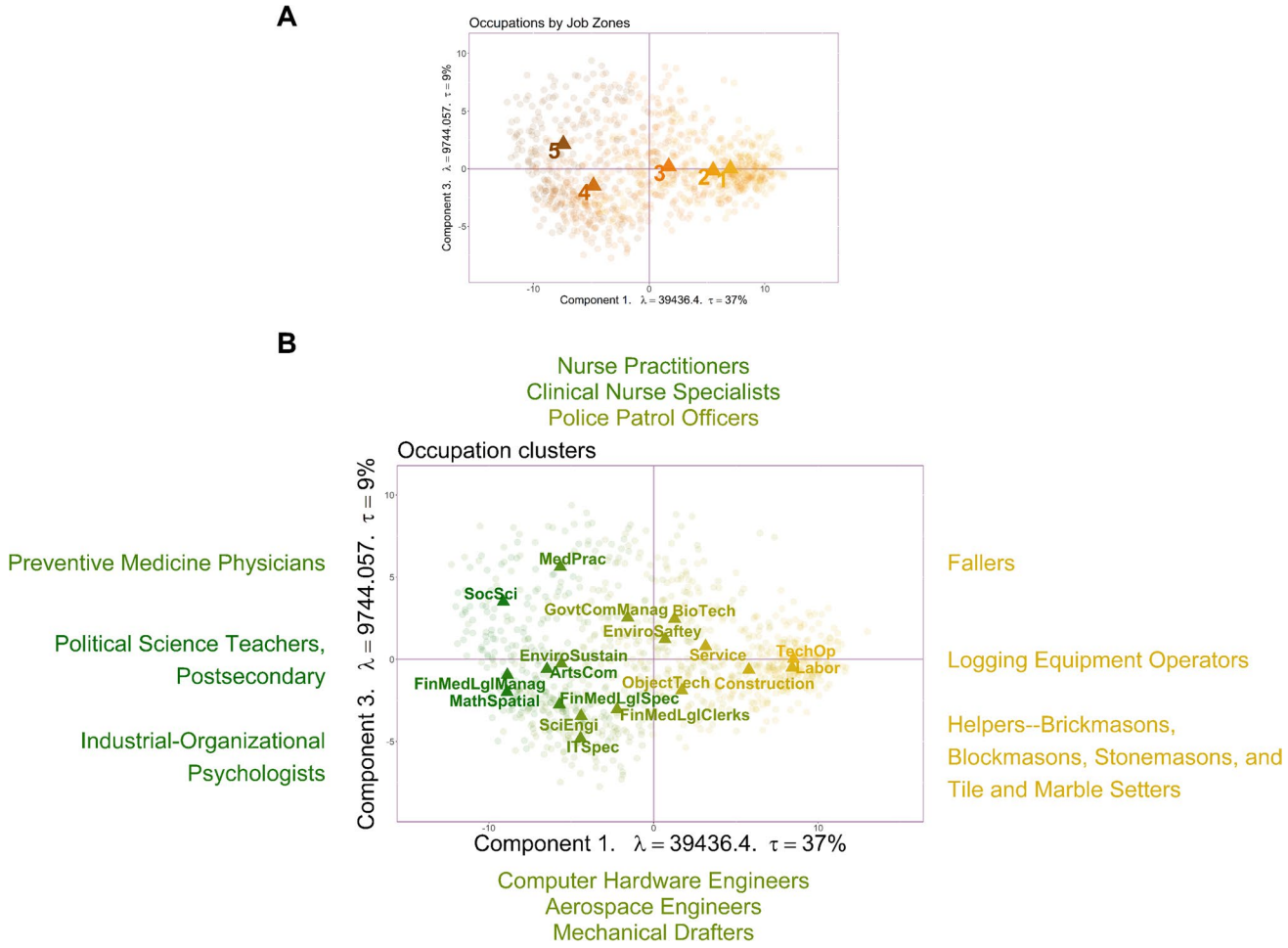

Fig. 3 First and third components (Preparation and Health versus Computational Science) for the General occupation PCA. *Note.* Panel **A** illustrates the degree to which each component corresponds to mean Job Zone ratings. Anchors indicate occupations (**B**) with the highest contributions (e.g., [Tree] Fallers and extent flexibility contributed strongly and positively to Preparation [Component 1]). In all plots, the horizonal axis represents Preparation (Component 1), and the vertical axis represents Health versus Computational Science (Component 3)

Supplementary S7, horizontal axes). Unlike the general occupation PCA, where the social science, humanities, and communication traits have scaled loadings close to null, these traits strongly contributed to the variance of this component. This difference is attributable to the greater influence of liberal arts education in this set of professions. We labeled this component "STEM versus Social Science and Humanities."

The second component of Job Zones 4–5 PCA, explaining 16% of the variance, differentiated occupations in health sciences and medicine (e.g., surgeons, nurse practitioners, oral and maxillofacial surgeons) from occupations in computer sciences and business (e.g., data specialists, cost estimators, research analysts), with the scaled loadings of traits reflecting social science, humanities, and natural science versus science, engineering, and math (Fig. 4B and Supplementary S7, vertical axes). We labeled this component "Health versus Computational Science."

Job zones 1–3 PCA

Job Zones 1–3 PCA identified seven significant components, which, together, explained 75% of the total variance. All significant components were included in the HCA, which identified ten occupation clusters and nine trait clusters (Fig. 1C and Supplementary S4). Job Zones 1–3 typically contain occupations that are labor-intensive (e.g., manual labor, typing speed).

Figure 5 shows the occupation space of the first two components of this PCA. The first component, explaining 33% of the total variance differentiated manual labor occupations (e.g., manufactured building installers, mechanics, millwrights) from office administrative occupations (e.g., telemarketers, clerks), with the scaled loadings of traits reflecting engineering, technology, operations and control, and physical strength as distinct from communication and humanities (i.e., customer-service-related skills; Fig. 5B and Supplementary

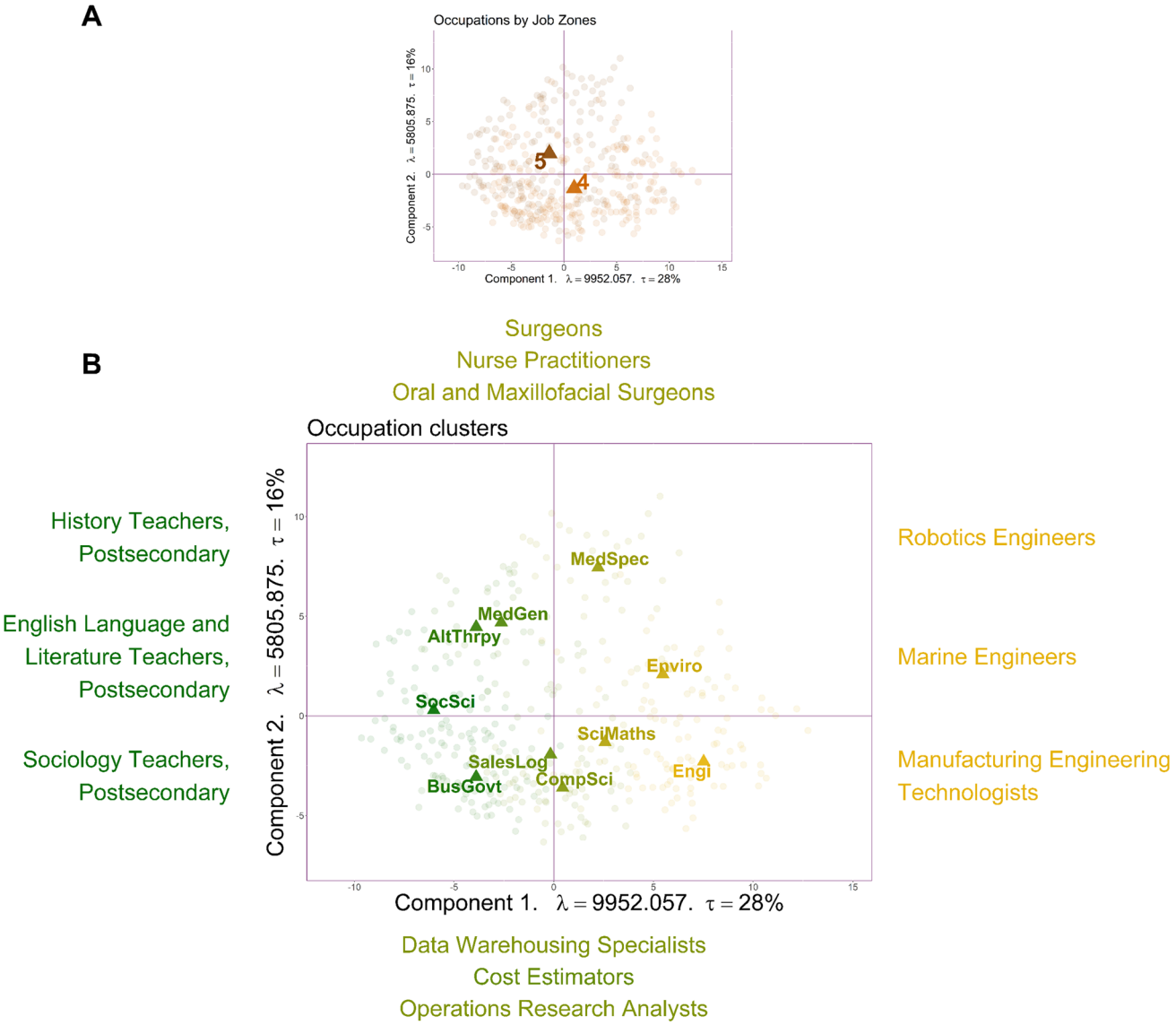

Fig. 4 First and second components (STEM versus Social Science and Humanities; Health versus Computational Science) for the Job Zones 4–5 PCA. *Note.* Panel **A** illustrates the degree to which each component corresponds to mean Job Zone ratings. Anchors indicate occupations (**B**) with the highest contributions (e.g., Engineers and Teachers contributed strongly to STEM versus Social Science and Humanities [Component 1]). Occupation component scores (**B**) are labeled by clusters derived from the HCA. In all plots, the *horizonal axis* represents STEM versus Social Science and Humanities (Component 1) and the *vertical axis* represents Health versus Computational Science (Component 2)

S8, horizontal axes). This component was labeled "Manual Labor versus Office."

The second component, explaining 19% of the variance, differentiated occupations that require specialized and practical knowledge and skills (i.e., technical occupations, such as product managers or fire-fighting supervisors) from occupations that do not require specialized and practical knowledge and skills (e.g., pressers, graders, sorters, cleaners). The scaled loadings of traits reflect technical and scientific knowledge and communication (which requires practical training and textbook learning) as opposed to general physical abilities (which rely on broad motor capacity) (Fig. 5B and Supplementary S8, vertical axes). This component was labeled "Technical."

Application: Relationship of verbal and non-verbal abilities to STEM occupations

To illustrate an application of VOLCANO on real-world data, we analyzed the relationship of verbal and non-verbal abilities in relation to occupation in the Nathan Kline Institute-Rockland Sample (NKI-RS; Nooner et al., 2012), a data set that

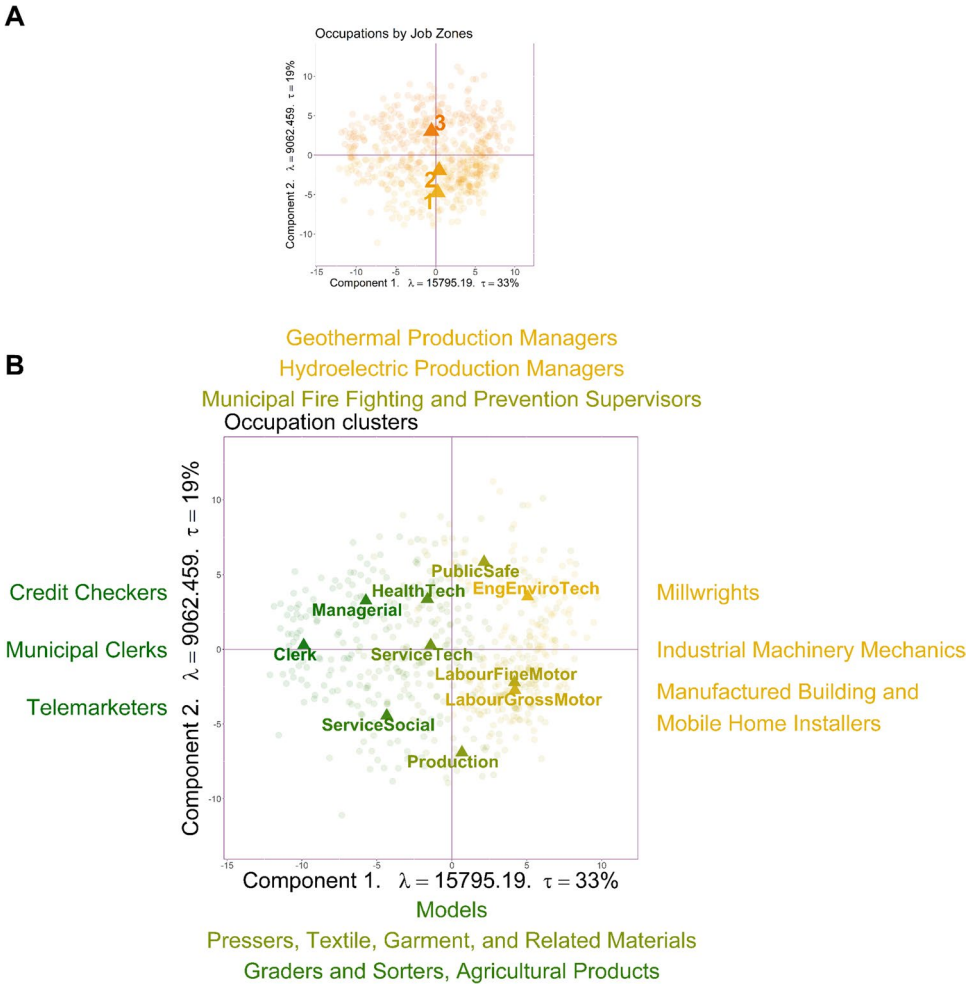

Fig. 5 First and second components (Manual Labor versus Office; Technical) for the Job Zones 1–3 PCA. *Note.* Panel **A** illustrates the degree to which each component corresponds to mean Job Zone ratings. Anchors indicate occupations (**B**) with the highest contributions (e.g., Mechanical occupations contributed strongly to Manual Labor versus Office [Component 1]). Occupation component scores (**B**) are labeled by clusters derived from the HCA. In all plots, the horizonal axis represents Manual Labor versus Office (Component 1) and the vertical axis represents Technical (Component 2)

has the advantage of rich standardized testing in a large sample of adults with occupation coded. As a proof-of-principle, we tested the hypothesis that non-verbal abilities would be uniquely associated with STEM occupations. We first created groups according to cognitive performance (Verbal/Non-verbal IQ discrepancy). For descriptive purposes, we examined frequencies of individuals in each discrepancy group within occupational clusters from our hierarchical clustering algorithm. We then projected groups defined by cognitive abilities into the occupational space to assess their relationships to the components. As the NKI-RS data focus on cognitive performance, we restricted this analysis to the Job Zones 4–5 PCA that includes occupations requiring more extensive training and preparation, especially with respect to cognitive skills. Indeed, the first two components of this space separate (1) STEM from Humanities and (2) Health from Computational Science. Finally, we extracted the occupation factor scores as continuous measures for relation to cognitive performance in a traditional regression analysis.

Methods

Because this analysis uses the Job Zones 4–5 PCA, only individuals who reported having an occupation that falls within Job Zones 4–5 were included in the analyses. A total of 470 NKI-RS participants were included (152 males and 318 females; $M_{age} = 56.69$ years, $SD_{age} = 15.45$ years). The NKI-RS dataset is a community sample of participants across the lifespan from Rockland County, a suburban/rural county 20 miles northwest of New York City. This sample is intended to be a phenotypically rich neuroimaging sample, consisting of data obtained from representative individuals from the

community rather than comprised solely of university students, as is typical of neuroimaging datasets. Individuals with past or current reports of head injury, stroke, bipolar disorder, autism spectrum disorder, attention deficit hyperactivity disorder, Alzheimer's disease, epilepsy, and a full-scale IQ < 70 were excluded from analyses. We also excluded individuals who did not speak English as their native language, because their verbal IQ scores would be artificially lowered. Only participants with both free entry occupation and scores on the Wechsler Abbreviated Scale of Intelligence – II (WASI-II) were included. Participants provided informed consent in accordance with the NKI-RS research ethics boards.

Participants manually entered their occupations as a free-entry item as part of the Hollingshead Four-Factor Index of Socioeconomic Status (Hollingshead, 1975). Free-text entries that do not correspond to an O*NET-listed occupation were re-coded by three independent raters who completed training on 100 representative occupations (the coding manual is available on GitHub: https://github.com/juchiyu/Occupation PCAs). Each rater converted 264 occupations, including 67 overlapping occupations. Inter-rater reliability for occupation coding of occupations that did not already correspond to an O*NET listed occupation (based on row cluster classification, see above Methods) was acceptable (Fleiss' mappa = .66; (Cohen, 1960; Fleiss et al., 1969).

Participants completed the Wechsler Abbreviated Scale of Intelligence – 2nd edition (WASI-II) as part of a comprehensive test battery. The WASI-II is a brief intelligence test—with excellent reliability and validity (Irby & Floyd, 2013)—designed for individuals between the ages of 6 and 90 years. This instrument includes four subtests Vocabulary, Similarities, Block Design, and Matrix Reasoning. A Verbal Comprehension Index (VCI) can be derived from the respective age-corrected standardized scores on the Vocabulary and Similarities subtests and a Perceptual Reasoning Index (PRI) can be derived from the respective age-corrected standardized scores on the Matrix Reasoning and Block Design subtests. VCI scores reflect verbal abilities including abstract verbal reasoning ability, semantic knowledge and verbal comprehension and expression. PRI scores reflect non-verbal abilities including visuospatial processing, and abstract problem solving. These measures correspond to the full Wechsler Adult Intelligence Scale – IV VCI and PRI scores, with a mean of 100 and a standard deviation of 15. We predicted that these indices would be differentially associated with Component 1 (STEM versus Social Science and Humanities) identified in Job Zones 4–5 PCA, with PRI associated with STEM occupation and VCI associated with social science and humanities occupations.

Participants were assigned to one of two groups based on their VCI-PRI discrepancy scores that were computed as the difference between VCI and PRI. Reliabilities of discrepancy scores in adults range from .82 to .89, a value considered large enough to justify hypothesis generation (Ryan & Gontkovsky, 2021). Individuals with a ten-point discrepancy between VCI and PRI and VCI were included in the projection. In this analysis, we only consider two groups from the distribution tails of the individuals: VCI+ and PRI+. The VCI+ group ($N = 145$) includes individuals with a VCI score greater than their PRI by at least ten points and were considered to have relatively stronger verbal skills. Conversely, PRI+ group ($N = 240$) includes individuals with a PRI score greater than their VCI by at least ten points and who, therefore, were considered to have relatively stronger perceptual reasoning skills; the remaining 85 participants difference score was less than ten points. The VCI+ and PRI+ groups did not differ by age $t(193) = -0.91$, *ns,* or gender $\chi^2(1) = 0.73$, *ns*. Participants were then colored based on group membership and projected into the cognitive occupation space, using *supplementary projections* (a procedure also called *out of sample elements projections,* for details, see Abdi & Williams, 2010).

Results and discussion

Figure 6A shows the occupation clusters with the proportion of each group for the purpose of sample characterization. The VCI+ group—relative to the PRI+ group—contains a larger proportion of people with low scores on Component 1 (e.g., occupations in Social Sciences, Business and Government, Alternative Therapies). Conversely, the PRI+ group, relative to the VCI+ group, contains a larger proportion of individuals with high scores on Component 1 of the Job Zones 4–5 PCA (e.g., occupations in Science and Mathematics, Engineering, Computer, and Informatics).

The utility of VOLCANO for quantifying occupational data on a ratio scale is illustrated by Figs. 6B and 6C. In Fig. 6B, the participants are represented as their occupations in the Job Zones 4–5 PCA space. The results show that individuals with stronger PRI relative to VCI scores have occupations that are more positive (i.e., higher in STEM) on Component 1 of the Job Zones 4–5 PCA. As the 95% bootstrapped confidence intervals (ellipses) do not overlap, this difference is statistically significant. There was no difference on Component 2 (Health versus Computational Science).

Next, we leveraged the full range of VCI and PRI scores in a multiple linear regression analysis predicting individuals' Component 1 score extracted from Job Zones 4–5 PCA plus gender. A significant regression was found $F(3, 466) = 19.35$, $p < 0.001$ with an R^2 of .11. Within this model, there was a significant main effect of gender ($\beta = -1.53$, SE = 0.30, $t = -5.06$, $p < 0.001$), with males having jobs with higher scores on Component 1 (i.e., males were more likely than females to have jobs in a STEM discipline). There was also a significant main effect of VCI and PRI scores (VCI: $\beta = -0.04$, SE = 0.01, $t = -2.88$, $p < 0.005$; PRI $\beta = -0.06$, SE = 0.01, $t = 5.45$, $p < 0.001$). Notably, there was no significant interaction

between gender and VCI or PRI. Additional results for this linear regression are presented in Table 3 and illustrated in Fig. 6B, where it can be seen that PRI was significantly positively related to Component 1 scores, $r(468) = .211$, $p < 0.001$, whereas VCI had a negative slope that was not significant, $r(468) = -.01$, *ns*, with the slopes of these bivariate correlations significantly different as examined by William-Hotelling's test, $t(467) = 4.93$, $p < 0.001$. To further illustrate how verbal versus non-verbal cognitive abilities relate to Component 1, the correlation of the VCI minus PRI difference score was significant, $r(468) = -.24$, $p < 0.001$ (see Fig. 6D).

These results illustrate the utility of VOLCANO for characterizing occupation and testing predictions on real-world occupational data. The empirically-derived clusters are useful for sample characterization, but data analytic options are restricted to non-parametric statistics. Many studies of occupation use such measures (Zeman et al., 2020). The added value of VOLCANO is the derivation of component scores for use in more powerful parametric analyses. This exercise was intended as proof-of-principle rather than theory-testing, as it generally accepted that spatial reasoning is important for STEM disciplines (e.g., Khine, 2017), whereas there was no *a priori* reason for this dissociation to be observed when contrasting Health versus Computational science (Component 2). The relationship between specific verbal intellectual abilities and selection of occupations in the social sciences and humanities has received less attention, possibly owing to the heterogeneity of these occupations.

Discussion

We used multivariate methods (i.e., PCA and HCA) to convert the heterogeneous categorical variable "occupation" into a concise set of continuous variables (along with the constrained set of categorical groupings). Our VOLCANO Shiny app provides a platform for standardized, quantitative characterization of occupation, enabling a new level of data sharing and comparison across studies concerning the skills, abilities, and traits associated with specific occupations. In addition to making O*NET data accessible and shareable, the VOLCANO Shiny app makes data flexible and supporting researcher's ability to use O*NET data to address a range of research questions.

We implemented PCA on traits of occupations from O*NET and revealed three meaningful continuous components. These include (1) a component that reflects the education and preparation needed for specific occupations, (2) a STEM component that reflects the degree to which occupations are within a STEM discipline, and (3) a component that distinguishes STEM occupations between those in health science and those from scientific professions that require computational and mathematical thinking. These components are similar to those previously described for DOT and O*NET data (Hadden et al., 2004). However, we seek to transcend a static occupational space to create a flexible, accessible application that can accommodate the dynamic needs of researchers studying occupational traits across disciplines.

The general occupation space derived in the current study is useful for questions associated with occupations across all Job Zones. The inclusion of two additional occupation spaces (i.e., Job Zones 4–5 and Job Zones 1–3) is useful for questions tailored to specific groups, such as higher cognitive skills (e.g., working memory, mathematical reasoning) for Job Zones 4–5 space or physical skills (e.g., basal metabolic rate) for the Job Zones 1–3 space. Together, these three occupation spaces hold promise to uncover how occupation is associated with a wide set of health, psychological, and financial measures as well as overall standard of living.

The first and second components of the Job Zones 4–5 occupation space closely resembled the second and third components of the general occupation space (i.e., STEM component and Health versus Computational Science component). This finding was expected because the components of PCA are orthogonal (i.e., the second and third components in the general occupation PCA are uncorrelated with the first education-related component). More importantly, removing the education-related component allowed for a wider distribution of cognitive skills along Components 2 and 3, enabling finer-grained distinctions across occupations requiring higher education, that did not contribute to the corresponding component in the general occupation PCA. By contrast, the Job Zones 1–3 occupation space—which included labor-intensive occupations—has distinct components. The first component separated manual labor from office jobs, while the second component separated occupations that rely more on technical skills from those that rely less on such skills. It is worth noting that this pattern is related to a linear pattern of Job Zones 1–3 (see Fig. 5A) which indicates that the education and the preparation required for these occupations are closely related to their technicality. The HCA, performed on the scaled loadings from the PCAs, provided a set of data-driven categories that can be used to address research questions that require categorical clusters of occupational traits.

Notably, across PCAs the heterogeneity of one component is often decomposed by the subsequent component. For instance, the heterogeneous construct of an occupation in a STEM discipline is often decomposed by the subsequent component. More specifically, Component 3 of the general PCA and Component 2 of the Job Zones 4–5 PCA; are labeled as "Health versus Computational Science." Because PCA components are orthogonal (i.e., uncorrelated)), these results showed that the distinction between non-STEM versus STEM (Health and Computational Science combined) is independent from the distinction within STEM (Health vs. Computational). Thus, our method does capture the complexity of professions such as medical doctors, who score high on health and low on computational science.

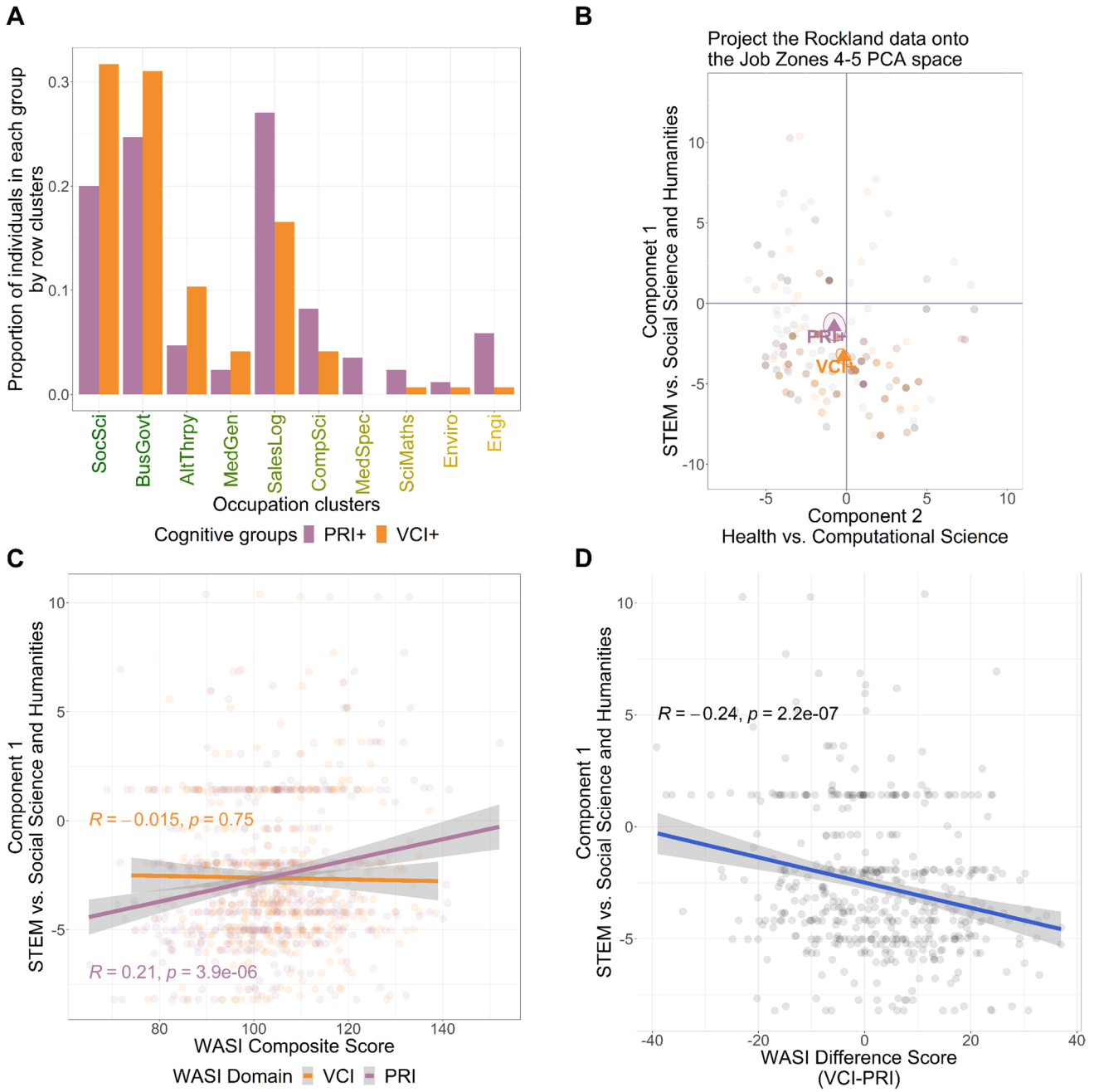

Fig. 6 Projection the Rockland data onto the space of Job Zones 4–5 PCA. *Note*. Visualization of the association between the Job Zones 4–5 PCA and the Wechsler Abbreviated Scale of Intelligence (WASI-II) including verbal comprehension index (VCI) and perceptual reasoning index (PRI) from the NKI-RS dataset. **A** shows the distribution of occupation clusters within the groups defined as VCI+ and PRI+. From the left, clusters range from low to high on Component 1 (see Figs. 1B and 4B for interpretation). As expected, the VCI+ group contains a larger proportion of people with occupations low on Component 1 (Social Sciences, Business and Government, Alternative Therapies) relative to the PRI+ group, whereas the PRI+ group contains more people with occupations high on Component 1 (Science and Mathematics, Engineering, Computer and Informatics). **B** Illustrates the supplementary projection of the discrepancy scores (VCI+ denotes the group with VCI larger than PRI by 10, and PRI+ denotes the group with PRI larger than VCI by 10). The *ellipses* indicate 95% bootstrapped confidence intervals. **C** The PRI is positively correlated with Component 1 scores, indicating an association with STEM, whereas the VCI is negatively correlated with Component 1, indicating an association with social sciences and humanities. The two correlation coefficients are significantly different. **D** The difference between participants' VCI and PRI scores (i.e., VCI-PRI) is negatively correlated with Component 1, indicating that participants with a smaller difference between VCI and PRI are more likely to have occupations in Social Science and Humanities than in STEM. For positive WASI difference scores, a greater difference between VCI and PRI relates to having a lower score on Component 1 (i.e., an occupation in the humanities). Conversely, for negative WASI difference scores, a greater difference between VCI and PRI relates to having a higher score on Component 1 (i.e., an occupation in STEM)

Table 3 Multiple linear regression results

	β	SE	t	p	Partial η^2
Gender	– 1.52	0.30	– 5.06	< 0.001	.06
VCI	– 0.04	0.01	– 2.88	< 0.005	.0002
PRI	0.06	0.01	5.45	< 0.001	.06

Multiple linear regression model predicting the cognitive occupational factor scores (Component 1) from verbal comprehension index (VCI) and perceptual reasoning index (PRI) of the Wechsler Abbreviated Scale of Intelligence (WASI-II)

The occupation spaces can be exploited by projecting new observations onto the components using supplementary projections. For instance, groups with specific exposures, neurological characteristics, or cognitive traits can be projected into the occupational space to determine the association between these supplementary traits and occupation selection (provided they have been coded within the same O*Net job titles that we used to create the spaces). Alternatively, researchers can use the Shiny app to extract occupations, component scores, scaled loadings, and categorical clusters for use in classic univariate analyses as well as complex multilevel modelling, such as structural equation modelling (SEM).

As a practical illustration of these methods, we assessed the relationships of verbal and non-verbal intellectual abilities to STEM versus social science / humanities occupations (Component 1 of Job Zones 4–5 PCA). Using both supplementary projections and extraction of component scores, we found that visuospatial and non-verbal analytical reasoning abilities were related to selection of STEM professions, as expected on the basis of prior research (Khine, 2017). Vocabulary and verbal reasoning were related to the practice of professions in the social sciences and humanities. Because these analyses were focused on the association between cognition and occupation, we restricted data to occupations that require extensive preparation and training and that are relatively more reliant on cognition (i.e., occupation space as defined by the Job Zones 4–5 PCA). A more extensive cognitive battery would be required to assess the association between other theoretically relevant skills and occupations across the full range of job zones. These findings were not intended to advance theory, but rather to provide proof-of-principle that the VOLCANO technique can be used to isolate specific occupational components in relation to external measures. Considering the NKI-RS sample, VOLCANO enables the incorporation of quantified occupational data into analyses with deep behavioral, mental health, and neuroimaging data included with that dataset. Moreover, VOLCANO standardization can facilitate linkages of findings across datasets containing O*NET-coded occupations.

We share our code on OSF and GitHub, and we provide a Shiny app to support researchers repurposing these data to address distinct research questions. With the publicly available code, the occupation space can be re-generated with an updated O*NET database. The Shiny app can be used to implement different clustering methods (either hierarchical clustering analysis or *K*-means), select different numbers of clusters, and generate distinct component spaces based on the inclusion of specific Job Zones (e.g., an occupational space within a single job zone). The Shiny app can also generate detailed lists of occupations and traits that comprise each component and cluster to help researchers identify key characteristics of the component and facilitate generating study-appropriate labels and names.

The following limitations should be considered by researchers using these methods. Each occupation's contribution within a component is a relative measure only interpretable within the full set of occupations included in the space. For instance, the same occupation would appear to be more physically demanding in the cognitive compared to the labor-intensive occupation space. Second, the occupational spaces are limited to those occupations listed in O*NET. While other out-of-sample occupations evaluated by the same set of traits could be projected onto the same space using supplementary projections (Abdi & Williams, 2010), such an exercise would assume that new trait ratings are comparable to older ones—an assumption that may be unjustified depending on the methods used to collect the ratings. Additionally, the definition of Job Zone is somewhat ambiguous because it reflects to varying degrees the education, training, experience, required knowledge, wage and salary level associated with particular jobs. That said, O*NET is the most comprehensive occupation dataset available. We acknowledge that routine updating of O*NET could change occupational spaces. We therefore provide the code required to generate occupational spaces using any occupation dataset, including future iterations of O*NET. Finally, we used consensus to derive labels for components and clusters in relation to the underlying constructs, but this labelling method is ultimately subjective.

Research implementing component reduction methodologies to characterize occupation data date back over 50 years, even before the emergence of comprehensive taxonomic systems, such as O*NET (e.g., Cole et al., 1971; Cunningham et al., 1983). Yet none of the prior attempts to reduce O*NET data has resulted in an accessible, sharable, and standardized system for the quantification and classification of occupation characteristics. The occupational space derived in the current study results in a set of components that converge with past research (e.g., Potter et al., 2008; Smyth et al., 2004; Spreng et al., 2010), but with additional intricacy, flexibility, and specificity. Moreover, this occupational space can be easily implemented, reproduced, updated, and adapted across disciplines and countries to enhance feasibility, consistency, and comparability across research settings. A key contribution of the current study is that it provides a flexible and easily accessible tool that will support and expedite future research on cognitive, neurological, and behavioral characteristics associated with occupations across a range of educational attainment levels. Additionally, findings from the

current study provide—for occupational traits and factors (i.e., components)—conceptual and terminology guardrails that will improve replicability and communication of findings across disciplines. Finally, our findings and techniques support research questions that will deepen our understanding of occupation, which in turn holds promise to enhance individual quality of life, and global innovation and productivity.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-022-02044-7.

Acknowledgements This work was supported by operating grants from the Canadian Institutes of Health Research (CIHR) (Grant no. MOP-148940), the Social Sciences and Humanities Research Council (SSHRC) (Grant no. 430-2020-00215), as well as an SSHRC Banting Post-Doctoral Fellowship to H.M.S.

Funding This work was supported by operating grants from the Canadian Institutes of Health Research (CIHR) (Grant no. MOP-148940), the Social Sciences and Humanities Research Council (SSHRC) (Grant no. 430-2020-00215), as well as an SSHRC Banting Post-Doctoral Fellowship to H.M.S.

Data availability The data that support the findings of this study are available at https://github.com/juchiyu/OccupationPCAs. These data were derived from the following resources available in the public domain: https://www.onetonline.org/. All analyses, data, and clustering are available through the Open Science Framework (OSF), GitHub (https://github.com/juchiyu/OccupationPCAs), and an original Shiny app (the codes are included in the GitHub and the link will be available after the anonymous review process).

Declarations

Ethic approval The occupation data is from a public data set and does not involve human subjects or animals and thus does not require ethical approval. Participants included in the application section provided informed consent in accordance research ethics boards.

Conflict of interest All authors declare that they have no conflicts of interest.

Permission to reproduce materials from other sources This study used data from Occupational Classification Network (O*NET) database (https://www.onetonline.org/) and US Department of Labor, Employment and Training Administration (https://www.doleta.gov/). The license that granted the permission to reproduce materials from O*NET is an Attribution 4.0 International license (https://creativecommons.org/licenses/by/4.0/), and no changes were made to the original data.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433–459. https://doi.org/10.1002/wics.101

Beaton, D., Fatt, C. R. C., & Abdi, H. (2014). An ExPosition of multivariate analysis with the singular value decomposition in R. *Computational Statistics & Data Analysis, 72*(0), 176–189. https://doi.org/10.1016/j.csda.2013.11.006

Bridges, C. C. (1966). Hierarchical cluster analysis. *Psychological Reports, 18*(3), 851–854. https://doi.org/10.2466/pr0.1966.18.3.851

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research, 27*(4), 509–540. https://doi.org/10.1207/s15327906mbr2704_2

Burrus, J., Jackson, T., Xi, N., & Steinberg, J. (2013). Identifying the MOST important 21ST century WORKFORCE competencies: An analysis of the occupational information network (O*Net). *2013*(2), i-55. https://doi.org/10.1002/j.2333-8504.2013.tb02328.x

Burzynska, A. Z., Jiao, Y., & Ganster, D. C. (2019). Adult-life occupational exposures: Enriched environment or a stressor for the aging brain? *Work, Aging and Retirement, 5*(1), 3–23. https://doi.org/10.1093/workar/way007

Clark, C. L. (2002). *Factor structures of the O* NET occupational descriptors* [Master's thesis, North Carolina State University]. Retrieved July 5, 2022, from http://www.lib.ncsu.edu/resolver/1840.16/640.

Cohen, J. (1960). *A Coefficient of Agreement for Nominal Scales., 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cole, N. S., Whitney, D. R., & Holland, J. L. (1971). A spatial configuration of occupations. *Journal of Vocational Behavior, 1*(1), 1–9. https://doi.org/10.1016/0001-8791(71)90002-9

Crouter, A. C., Lanza, S. T., Pirretti, A., Goodman, W. B., & Neebe, E. (2006). The O*Net jobs classification system: A primer for family researchers. *Family Relations, 55*(4), 461–472. https://doi.org/10.1111/j.1741-3729.2006.00415.x

Cunningham, J. W., Boese, R. R., Neeb, R. W., & Pass, J. J. (1983). Systematically derived work dimensions: Factor analyses of the occupation analysis inventory. *Journal of Applied Psychology, 68*(2), 232–252. https://doi.org/10.1037/0021-9010.68.2.232

de Vries, A., & Ripley, B. D. (2020). Ggdendro: Create Dendrograms and tree diagrams using 'ggplot2'. Retrieved May 3, 2022, from https://CRAN.R-project.org/package=ggdendro.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*(3), 211–218. https://doi.org/10.1007/BF02288367

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*(5), 323–327. https://doi.org/10.1037/h0028106

Fouad, N. A., & Kozlowski, M. B. (2019). Turning around to look ahead: Views of vocational psychology in 2001 and 2019. *Journal of Career Assessment, 27*(3), 375–390. https://doi.org/10.1177/1069072719841602

Gadermann, A. M., Heeringa, S. G., Stein, M. B., Robert Ursano, C. J., Lisa Colpe, C. J., Fullerton, C. S., . . . Kessler, R. C. (2014). Classifying U.S. Army military occupational specialties using the occupational information network. *Military Medicine*, *179*(7), 752–761. https://doi.org/10.7205/MILMED-D-13-00446/J

Galili, T. (2015). Dendextend: An R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics, 31*(22), 3718–20. https://doi.org/10.1093/bioinformatics/btv428

Hadden, W. C., Kravets, N., & Muntaner, C. (2004). Descriptive dimensions of US occupations with data from the O*NET. *Social Science Research, 33*(1), 64–78. https://doi.org/10.1016/S0049-089X(03)00039-5

Handel, M. J. (2016). The O*NET content model: Strengths and limitations. *Journal for Labour Market Research, 49*(2), 157–176. https://doi.org/10.1007/s12651-016-0199-8

Hanson, M. A., Borman, W. C., Kubisiak, U. C., & Sager, C. E. (1999). *Cross-domain analyses*.

Hartman, R. O., & Betz, N. E. (2007). The five-factor model and career self-efficacy: General and domain-specific relationships. *Journal of Career Assessment, 15*(2), 145–161. https://doi.org/10.1177/1069072706298011

Habeck, C., Eich, T. S., & Stern, Y. (2019). Occupational patterns of structural brain health: Independent contributions beyond education, gender, intelligence, and age. *Frontiers in Human Neuroscience, 13*, 1–7. https://doi.org/10.3389/fnhum.2019.00449

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Psychological Assessment Resources.

Hollingshead, A. B. (1975). *Four factor index of social status. In*. New Haven, CT.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417–441. https://doi.org/10.1037/h0071325

Indiana Department of Workforce Development Research and Analysis Division, & Indiana Business Research Center. (2011). *Navigating change: Exploring new Career pathways in an evolving labor market*. Retrieved October 26, 2022, from http://www.drivingworkforcechange.org/reports/careerpathways.pdf

Irby, S. M., & Floyd, R. G. (2013). Review of Wechsler abbreviated scale of intelligence, second edition. *Canadian Journal of School Psychology, 28*, 295–299. https://doi.org/10.1177/0829573513493982

Khine, M. S. (2017). Spatial cognition: Key to STEM success. In *Visual-spatial ability in STEM education* (pp. 3–8). Springer.

Larson, L. M., Rottinghaus, P. J., & Borgen, F. H. (2002). Meta-analyses of big six interests and big five personality factors. *Journal of Vocational Behavior, 61*(2), 217–239. https://doi.org/10.1006/jvbe.2001.1854

Levine, J. D. (2003). *Use of the O* NET descriptors in numerical occupational classification: An exploratory study (publication number 3098975) [doctoral dissertation, North Carolina State University]*. ProQuest Dissertations Publishing.

Nolan, C., Morrison, E., Kumar, I., Galloway, H., & Cordes, S. (2011). Linking industry and occupation clusters in regional economic development. *Economic Development Quarterly, 25*(1), 26–35. https://doi.org/10.1177/0891242410386781

Nooner, K., Colcombe, S., Tobe, R., Mennes, M., Benedict, M., Moreno, A., & Milham, M. (2012). The NKI-Rockland sample: A model for accelerating the pace of discovery. *Science in Psychiatry [Review]., 6*. https://doi.org/10.3389/fnins.2012.00152

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., & Dye, D. M. (2001). Understanding work using the occupational information network (O*NET): Implications for practice and research. *Personnel Psychology, 54*(2), 451–492. https://doi.org/10.1111/j.1744-6570.2001.tb00100.x

Potter, G. G., Helms, M. J., & Plassman, B. L. (2008). Associations of job demands and intelligence with cognitive performance among men in late life. *Neurology, 70*(19 Part 2), 1803–1808. https://doi.org/10.1212/01.wnl.0000295506.58497.7e

R core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved October 20, 2022, from https://www.R-project.org/

Ralston, C. A., Borgen, F. H., Rottinghaus, P. J., & Donnay, D. A. C. (2004). Specificity in interest measurement: Basic interest scales and major field of study. *Journal of Vocational Behavior, 65*(2), 203–216. https://doi.org/10.1016/S0001-8791(03)00096-4

Reddon, J. R. (1984). *The number of principal components problem: A Monte Carlo study* (publication number 1379) [doctoral dissertation, the University of Western Ontario]. Retrieved July 5, 2022, from https://ir.lib.uwo.ca/digitizedtheses/1379

Ryan, J. J., & Gontkovsky, S. T. (2021). *Reliabilities of Discrepancy Scores and Supplemental Tables for the WASI–II., 39*(8), 930–937. https://doi.org/10.1177/07342829211040595

Savickas, M. L. (2001). The next decade in vocational psychology: Mission and objectives. *Journal of Vocational Behavior, 59*(2), 284–290. https://doi.org/10.1006/jvbe.2001.1834

Shu, X., Fan, P.-L., Li, X., & Marini, M. M. (1996). Characterizing occupations with data from the dictionary of occupational titles. *Social Science Research, 25*(2), 149–173. https://doi.org/10.1006/ssre.1996.0007

Slaper, T. F. (2014). Clustering occupations. *Indiana Business Review, 89*(2), 7–12. Retrieved July 5, 2022, from https://www.ibrc.indiana.edu/ibr/2014/summer/article2.html

Smart, E. L., Gow, A. J., & Deary, I. J. (2014). Occupational complexity and lifetime cognitive abilities. *Neurology, 83*(24), 2285–2291. https://doi.org/10.1212/wnl.0000000000001075

Smyth, K. A., Fritsch, T., Cook, T. B., McClendon, M. J., Santillan, C. E., & Friedland, R. P. (2004). Worker functions and traits associated with occupations and the development of AD. *Neurology, 63*(3), 498–503. https://doi.org/10.1212/01.WNL.0000133007.87028.09

Spreng, R. N., Rosen, H. J., Strother, S., Chow, T. W., Diehl-Schmid, J., Freedman, M., & Levine, B. (2010). Occupation attributes relate to location of atrophy in frontotemporal lobar degeneration. *Neuropsychologia, 48*(12), 3634–3641. https://doi.org/10.1016/j.neuropsychologia.2010.08.020

United States Department of Labor, U. s. E. o. S., North Carolina Occupational Analysis. (2006). *Dictionary of Occupational Titles (DOT)* (Rev. 4th ed.). 10.3886/ICPSR06100.v1

US Department of Labor. (2019a). *O*NET-SOC Taxonomy at O*NET Resource Center.* O*NET OnLine. Retrieved October 27, 2022, from https://www.onetcenter.org/taxonomy.html

US Department of Labor. (2019b). *The O*NET Content Model*. O*NET OnLine. Retrieved October 17, 2022, from https://www.onetcenter.org/content.html

US Department of Labor. (2019c). *O*NET OnLine Help: scales, ratings, and standardized scores*. O*NET OnLine. Retrieved July 5, 2022, from https://www.onetonline.org/help/online/scales#score

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. Retrieved October 20, 2022, from https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of open source software, 4*(43), 1686.

Zeman, A., Milton, F., Della Sala, S., Dewar, M., Frayling, T., Gaddum, J., & Winlove, C. (2020). Phantasia-the psychological significance of lifelong visual imagery vividness extremes. *Cortex, 130*, 426–440. https://doi.org/10.1016/j.cortex.2020.04.003

Open practices statement The data and materials for all analyses in this study are available at https://github.com/juchiyu/OccupationPCAs, with the exception of the proof-of-principal data as a data use agreement (DUA) is required to access the phenotypic data from the Rockland Sample. Information on how to access this data can be found at data.rocklandsample.rfmh.org.

Data transparency table The data reported in this manuscript were obtained from publicly available data, O*NET, https://www.onetcenter.org/overview.html. A bibliography of journal articles, working papers, conference presentations, and dissertations using the O*NET is available at https://www.onetcenter.org/references.html. The variables and relationships examined in the present article have not been examined in any previous or current articles, or to the best of our knowledge in any papers that will be under review soon.