2022 No. 036

Developing Related Occupations for the O*NET Program

Prepared for: National Center for O*NET Development
313 Chapanoke Road, Suite 130
Raleigh, NC 27603

Prepared under: Subcontract Number
(through RTI International):
1-312-0207142--41224L

Authors: Jeffrey A. Dahlke, HumRRO
Dan J. Putka, HumRRO
Ori Shewach, HumRRO
Phil Lewis, National Center for O*NET Development

Date: April 12, 2022

Developing Related Occupations Information for the O*NET Program

Table of Contents

Table of Contents

List of Tables

List of Figures

Developing Related Occupations Information for the O*NET Program

Introduction

The Occupational Information Network (O*NET) is a comprehensive system developed by the U.S. Department of Labor that provides information on over 900 occupations within the U.S. economy. This information is maintained in a comprehensive database (at the time of this report, the most current version is the O*NET 26.2 Database; National Center for O*NET Development, 2022). To keep the database current, the National Center for O*NET Development (hereafter referred to as "the Center") is involved in a continual data collection process aimed at identifying and maintaining current information on the characteristics of workers and occupations. For years, the Human Resources Research Organization (HumRRO) has supported the Center's efforts to maintain the database. The purpose of this proposed work was to develop and evaluate a new process for identifying "related occupations" for the 923 data-level occupations included within the O*NET-SOC 2019 taxonomy (Gregory et al., 2019).

The Center has a long-standing history of conducting research to populate related occupations information for the occupations included in its occupational taxonomy (Allen et al., 2012; Drewes et al., 1999). An individual's ability to discover and review related occupations, historically defined within O*NET as occupations with common work attributes and similar worker requirements, is a central component of both career exploration and job search/transition. Related occupations information is also often leveraged in a variety of other important world-of-work related activities, including human resource functions, workforce development, and basic research.

Current Related Occupations within the O*NET System

At the time of this research, related occupations information within the O*NET System is available for 777 of the 923 data-level occupations. There is not full coverage of occupations due to (a) missing input data required during the current data's development effort and (b) the transition to a new occupation classification structure (e.g., new or modified occupations added; occupations removed).

Two sets of related occupations are currently available:

- Career Changers Matrix: Related occupations that make use of similar skills and experience; workers from one occupation may transfer to a job in a related occupation with minimal additional preparation.
- Career Starters Matrix: Related occupations that make use of similar general capabilities and interests; career explorers interested in the reference occupation may also be interested in the related occupations.

For both matrices, each O*NET-SOC occupation has up to ten rank-ordered related O*NET-SOC occupations. Rank is based on the goal of the particular the matrix (described above).

Related occupation information is incorporated within occupation/career reports in the O*NET websites:

- O*NET OnLine and O*NET Code Connector use the Career Changer matrix and display up to 10 occupations in order of O*NET-SOC codes.
- My Next Move, My Next Move for Veterans, and Mi Proximo Paso use the Career Starter matrix and display up to 5 occupations in alphabetical order.

In addition, both matrices are available to developers and researchers via the O*NET Database, O*NET OnLine Web Services, and O*NET My Next Move Web Services.

Project Goals

At the outset of this project, the Center communicated the following goals:

1) Develop related occupations based on the similarity of work performed that will be more intuitive for customers. Consider potentially moving to a single, universal matrix structure that can be filtered or adjusted based on customer/audience need, rather than two distinctly developed matrices.

2) Primarily use the subset of O*NET data consistently available for all data-level occupations, thus allowing for each data-level occupation to be populated with related occupations information, and for each data-level occupation to be an eligible candidate for other occupations' related occupations.

3) Determine if additional "supplemental" related occupations information beyond ten occupations can be developed and published.

4) Develop a method that is streamlined to facilitate future updates. Updates include maintaining the currency of the related occupation information via input data available from the O*NET data collection program (i.e., annual and quarterly updates). Also, updates involve developing related occupations information when a new O*NET-SOC taxonomy occupational classification structure is released (i.e., populating new-and-emerging occupations or significantly modified occupations).

The Current Project

The Center asked HumRRO to develop and evaluate a new process for identifying related occupations for the 923 data-level occupations included within the O*NET-SOC 2019 taxonomy. The Center expressed a desire to focus the determination of occupational relatedness on the subset of O*NET data currently available for the 923 datal-level occupations. Much of the available data is in the form of text, including occupation descriptions, task lists, and lists of work activities. This text-focused approach is quite different from the existing related occupations matrices available through O*NET, which are based heavily on level and importance data from a wide variety of O*NET cross-occupation domains (e.g., skills, knowledges, abilities, work styles, interests; Allen et al., 2012). The Center believed that its rich database of text about occupations would provide robust inputs for the development of strong, pragmatic related occupations linkages that will have utility and face validity for customers.

Overview of Technical Approach

At the outset of this work, the Center sought a method for identifying related occupations that was grounded in the notion of similarity of work performed. In other words, if the tasks and activities associated with two different occupations are similar, then it suggests the occupations are similar in substance and, in turn, "related." Individuals know what they do on their jobs, so framing relatedness in terms of the similarity of work performed on a job (as reflected in tasks and activities) is likely more straightforward for lay users and customers of O*NET to understand. Therefore, it has the potential to be accepted as more face valid compared to basing relatedness on more abstract worker-related characteristics (e.g., similarity of profiles that reflect knowledges, skills, abilities, or work styles). Beyond its potential value from a lay perspective, framing relatedness in terms of similarity of work performed is critical in the context of transporting validity evidence in the context of personnel decision making (Equal Employment Opportunity Commission et al., 1978; Society of Industrial and Organizational Psychology, 2018). Thus, beyond its potential clarity and value to lay users of O*NET, recasting related occupations in terms of similarity of work performed may also have value for technical users of O*NET who wish to leverage related occupations information for validity transportation efforts.

Given the Center's goals above, our initial approach was focused on identifying occupations that are related in terms of the work performed. Within O*NET, this information clearly manifests in the task statements, detailed work activities (DWAs), and occupation descriptions for each occupation. As such, we first sought to identify related occupations within O*NET based on the similarity of task statements, DWAs, and occupation descriptions, which we expressed in terms of a set of work-based occupational similarity metrics (WB-OSMs). We subsequently evaluated the validity of these metrics based on their relation to alternative indexes of occupational similarity that should theoretically flow from performing similar tasks and activities on the job (e.g., KSAO profile similarity, interest and work value profile similarity, salary similarity). Following this evaluation, HumRRO and the Center reviewed initial lists of related occupations that resulted from use of a composite of best-bet WB-OSMs identified in our initial evaluation.

Based on the review of the initial related occupations lists, HumRRO and the Center decided to adjust course and consider additional information that could help improve relatedness determinations compared to WB-OSMs alone. Specifically, we calculated and evaluated knowledge-based and alternate title-based similarity, in addition to work-based similarity, to determine occupational relatedness. Based on this evaluation, we decided to move forward with an occupational relatedness composite reflecting the following components: (a) task- and DWA-based similarity, (b) knowledge importance similarity, and (c) alternate title similarity. This final approach includes three important contributors to occupational similarity: what people in the occupations *do*, what they *know*, and what they are *called*.

We then constructed a new related occupations matrix based on this revised relatedness composite. We again reviewed a sampling of occupations, and the Center conducted a structured review of 60 representative occupations. Upon the Center's approval of the final relatedness composite, the Center reviewed the related occupations lists for all 923 data-level O*NET-SOCs. HumRRO then conducted a set of analyses to evaluate the quality of the resulting related occupation linkages. In the sections that follow, we detail each major phase of this work. We conclude this report with a description of the final work products that resulted from this effort and plans for their future updating.

Phase 1: Development and Evaluation of Initial Work-Based Occupational Similarity Metrics

For each pairing of the 923 data-level occupations included in the O*NET-SOC 2019 taxonomy, we began by calculating task, DWA, and occupation description work-based occupational similarity metrics (WB-OSMs) using the O*NET 25.3 Database (National Center for O*NET Development, 2021a).[1] For tasks, DWAs, and occupation descriptions, we calculated multiple types of similarity metrics. We outline each of these types below, starting with task-based similarity metrics.

Prior to calculating the similarity estimates described in the sections that follow, we took the following steps to prepare the text for analysis:

- We converted all text to lower case.
- We removed punctuation and separated hyphenated terms into individual words.
- For analyses that relied on word-level embeddings (described later), we identified substitute words for terms that were not included in the pre-trained embedding database (e.g., the database did not include the plural word "coatrooms," so we used the singular word "coatroom" in its place; the database did not include the word "batchmaking," so we used "batch making"). If there was not an acceptable substitute, we dropped the word from these analyses because it could not be represented quantitatively.[2]

Note that, when we used contextualized language models (described later), we did not apply any pre-processing to the text, as these models use punctuation cues as part of their process for quantifying features of text.

Task Statement Similarity

We calculated six types of task-based occupational similarity metrics that reflect different ways to represent each task as a numeric vector. For each type, we calculated two variants: one reflecting the simple, unweighted average of each task pair for the occupations being compared (i.e., tasks from occupation A compared with tasks from occupation B), and the other a weighted version that we describe further after introducing the five types below.[3] [4]

---

[1] In later of phases of this work, and for the final related occupations matrix delivered as a result of this project, we used the O*NET 26.1 Database (National Center for O*NET Development, 2021b)

[2] Out of 123 words that were not represented in the pre-trained GloVe database, we found suitable replacements for 106 words and dropped the remaining 17 words that had no analog in GloVe.

[3] When comparing occupations for similarity of tasks, we focused on core tasks and new tasks that have not yet been assigned a "Task Type" designation (i.e., we avoided analyzing tasks identified as supplemental: tasks that are less relevant and/important to the occupation). In the event an occupation didn't have ratings data for determining core task status (i.e., importance and relevance ratings), we used all task statements for that occupation.

[4] We conducted exploratory analyses that evaluate whether cosine similarity or angular distance (a simple transformation of cosine similarity) achieved better results. To date, cosine similarity metrics have dominated treatments of semantic similarity in the NLP literature, but the researchers who produced Google's Universal Sentence Encoder (Cer et al., 2018) found that angular distance can improve results by stretching out the top and bottom end of the cosine similarity distribution. For this reason, we viewed angular distance as offering the potential for more fine-grained distinctions of tasks similarity at the top of the distribution. We did not find that using angular distance improved our inferences, so we only present cosine-based results in this report.

- **Cos (TF)**: This estimates occupational similarity as the average of cosine similarity coefficients across all possible pairs of tasks between the occupations being compared, using term frequencies (TF; i.e., word counts) for all words that appear in a task as the vector representation for that task (Salton & Buckley, 1988). This method evaluates similarity in word usage across pieces of text, but it is literal and cannot account for synonyms or word context. For example, “the dog bit the man” and “the man bit the dog” would have a TF cosine of 1, while “businesses compensate workers” and “companies pay employees” would have a TF cosine of 0. Other methods we explored avoided this literal treatment of words and were able to account for synonyms, but we examined TF cosines because they provide a classic measure of sentence similarity.

- **Cos (TF-IDF)**: This estimates occupational similarity as the average of cosine similarity coefficients across all possible pairs of tasks between the occupations being compared, using the term frequency-inverse document frequencies (TF-IDF) for all words that appear in a task as the vector representation for that task (Salton & Buckley, 1988). Relative to TFs, TF-IDFs down-weight common words (e.g., the, a, an, and, or) when evaluating the similarity of any pair of tasks.[5] TF-IDF cosines suffer from the same general limitations as TF cosines because this method is literal and cannot account for synonyms or word context.

- **Cos (GloVe 297)**: This estimates occupational similarity as the average of cosine similarity coefficients across all possible pairs of tasks between the occupations being compared, using the average 297-element pre-trained GloVe (Global Vectors for Word Representation) word embeddings (Pennington et al., 2014; Common Crawl, 42B token database) as the vector representation for words that appear in a task.[6] Relative to TFs and TF-IDFs, GloVe embeddings provide a way to deal with words that are different in their text representations but similar in meaning (i.e., synonyms). For example, while “businesses compensate workers” and “companies pay employees” would have TF and TF-IDF cosines of 0, GloVe can account for word-level similarity and the GloVe-based cosine would be quite high. For each task, we computed the task-level embedding as the simple average of the task’s word-level embeddings for unique words used in the task statement. Using only the unique words within each task avoids inflating the influence of common words (e.g., the, a, an, and, or) while still getting a complete account of the task’s semantic content.

- **Cos (GloVe 297 TF-IDF)**: This estimates occupational similarity as the average of cosine similarity coefficients across all possible pairs of tasks between the occupations being compared, using the same embeddings as the Cos (GloVe 297) metric but using a different strategy to combine the word-level embeddings. For this metric, the vector representation for each task is a TF-IDF weighted average of the GloVe embeddings for each word in that task. This method has similar advantages to the Cos (GloVe 297) metric, with the additional advantage of down-weighting common words.

---

[5] When estimating IDFs for words, we based the IDFs on all unique sentences, core tasks, and new undesignated tasks (or all tasks for an occupation, when ratings to determine task importance were unavailable), and DWAs included in the O*NET 25.3 Database (National Center for O*NET Development, 2021a). Each sentence from an occupation description, each task, and each description was treated as a separate “document” for purposes of IDF calculation.

[6] The GloVe 42B token Common Crawl DB includes a 300-element vector for each word it contains, but three of those vectors are problematic because using them artificially inflates the cosines between words. Other researchers have independently identified the same issue (Lee et al., 2016). We omitted GloVe dimensions 7, 97, and 225 and based our analyses on the remaining 297 dimensions.

- **Cos (SBERT Embeddings)**: This estimates occupational similarity as the average of cosine similarity coefficients across all possible pairs of tasks between the occupations being compared, using sentence-level embeddings from SBERT (Sentence BERT; Reimers & Gurevych, 2019) as the vector representation of each task (via the "nli-distilroberta-base-v2" model). SBERT is built on top of the BERT architecture, which produces contextualized word-level embeddings. These embeddings are sensitive to the context in which a word is used. For example, in "I went to the bank" and "The bank of the river was beautiful," "bank" has a different word-level embedding in each sentence, while those embeddings would be identical in GloVe. SBERT then averages the word-level embeddings from BERT and uses the resulting sentence-level embedding as the input to another neural network that weights the sentence-level embedding to be optimized for detecting semantic similarity via cosines.
- **Cos (USE-DAN Embeddings)**: This estimates occupational similarity as the average of cosine similarity coefficients across all possible pairs of tasks between the occupations being compared, using sentence-level embeddings from Google's Universal Sentence Encoder (USE) with Deep Averaging Network (DAN) encoding (Cer et al., 2018) as the vector representation for each task. Like SBERT, USE-DAN accounts for differences in context between sentences. However, relative to other embeddings described above, USE-DAN offers true sentence-level embeddings rather than word-level embeddings that are averaged to form sentence-embeddings.

As alluded to above, for each similarity metric, we calculated *unweighted* estimates of the semantic similarity of task statements for each occupation pair, along with *weighted* estimates of semantic similarity, where the weights reflect the occupation-specific importance ratings tied to each task.[7] The potential value that weighted similarity estimates provide over their unweighted versions is that that they can factor in not only semantic similarity of the tasks being compared, but also the extent to which they are ordered similarly in the occupations being compared. For example, such weighting allows the similarity of two tasks that are highly important to the occupations being compared to carry more weight than the similarity of two tasks of lower importance when forming an overall estimate of task-based similarity for those occupations.

DWA Similarity

For DWAs, we calculated all six types of similarity metrics described above, using DWA statements for each pair of occupations being compared, rather than tasks statements (i.e., 12 similarity metrics in total: six types, with one set unweighted and the other set weighted). However, in the case of DWAs, the weighted variants on the metrics were computed differently. Specifically, the weights were a function of the rank-ordering of DWAs in terms of their importance to each occupation (National Center for O*NET Development, 2015).

[7] The majority of occupations had task importance ratings (i.e., 873). For occupations where no importance ratings are available, we equally weight tasks for that occupation. For occupations where there was a mix of tasks with ratings and without ratings, we used the average importance rating of tasks with ratings as proxy importance rating for tasks without ratings. For example, only a subset of tasks had importance ratings available for the 58 newly added data-collection-level occupations in the 2019 O*NET-SOC Taxonomy.

Occupation Description Similarity

For occupation descriptions, we first parsed each description into separate sentences (for occupations that had multiple-sentence descriptions). We then calculated all six types of similarity metrics above, using sentences from each occupation's description for each pair of occupations being compared as inputs rather than tasks or DWAs. Unlike tasks and DWAs, we only examined unweighted versions of these metrics since there are no ratings tied to individual sentences within O*NET occupation descriptions.

Summary

In total, we calculated 30 work-based occupational based similarity metrics: 12 that used tasks as input, 12 that used DWAs as input, and six that used sentences from occupation descriptions as input (please refer to Appendix A for a summary of how these WB-OSMs relate to each other). As noted above, the metrics range from simple word count-based indices to more complex language model-based indices and provided a diverse range of options to consider and evaluate.

Calculation of Occupational Similarity Criterion Variables

As noted earlier, if the work performed in a pair of occupations is similar (as manifested in the similarity of their tasks, DWAs, and/or descriptions), they should also exhibit similarity in other ways (e.g., similarity in salary and/or profiles of KSAs, interests, work values, or work styles). To address this possibility, we initially calculated the following estimates of occupational similarity for each pairing of the 923 data-level occupations included within the O*NET-SOC 2019 taxonomy for which requisite data were available in the O*NET 25.3 Database (National Center for O*NET Development, 2021a):

- Similarity with respect to ratings of O*NET knowledges, skills, abilities, interests, work values, and work styles (initially, separate similarity estimates for each domain)[8]
- Similarity with respect to median salary, obtained from the U.S. Bureau of Labor Statistics (2020)

For each O*NET domain above except for median salary, we calculated two similarity metrics: (a) the profile correlation of importance ratings across elements within the given domain (e.g., correlation of knowledge importance profile for occupation A and occupation B), and (b) squared Euclidean distance of importance ratings across elements within the given domain. Whereas the correlation frames similarity only in terms of shape (i.e., similarity in relative rank-ordering of knowledges for a pair of occupations), the latter frames similarity as a function of similarity in elevation (e.g., similar knowledge importance rating means) and scatter (e.g., similar knowledge importance rating standard deviations), as well as shape. For median salary, we calculated the absolute value of the median salary difference between occupation pairings, as we viewed magnitude—but not direction—of salary difference as conceptually meaningful for modeling purposes.

[8] We used importance ratings for knowledges, skills, abilities, and work styles when estimating similarity for the aforementioned domains. We used extent ratings for work values, and O*NET interest ratings reflecting how descriptive and characteristic an occupation is of each of the RIASEC interest dimension when estimating similarity for the work value and interest domains, respectively.

These calculations resulted in 13 potential criterion variables. After each was calculated, we examined distributions and intercorrelations among these variables across occupational pairs. We computed one final criterion variable (a composite of knowledge, skill, and ability distance scores) to reflect similarity in overall KSA profiles by elevation, scatter, and shape. We worked with the Center to consolidate this set of variables for use in our subsequent evaluation of the WB-OSMs.

Based on a review of data for these criterion variables (and conceptual considerations further outlined below), we decided to move forward with the following five criterion variables: a KSA distance composite, vocational interests shape similarity, work values shape similarity, work styles shape similarity, and absolute median salary difference. These criterion variables were selected to capture conceptually distinct aspects of occupational similarity across occupational KSAO profiles. We calculated distance scores as raw Euclidian distance, calculated as follows:

$$D = \sqrt{\sum (X_i - Y_i)^2}$$

where *X* and *Y* represent an occupation pairing and *i* represents individual knowledge, skill, and ability profile elements. Subsequently, we computed the KSA distance composite as:

$$KSA\ Distance\ Composite = Z\left(D_{Knowledges}\right) + Z(D_{Skills}) + Z(D_{Abilities})$$

where *Z* represents distance scores placed in standard score form (i.e., *z*-scores created by standardizing the given raw Euclidean distance across all occupation pairs). Shape scores were calculated as the Pearson correlation between occupational pairings on profiles of vocational interests, work values, and work styles, respectively. To calculate the absolute value of the median salary difference between occupations, we first imputed median salaries for four O*NET-SOCs with missing salary data, using hourly median wage as the sole predictor in a simple linear regression model. [9] Descriptive statistics and intercorrelations between the five selected criterion variables are presented in Table 1. As shown in this table, the five criteria were relatively independent of one another, with an average intercorrelation of *r* = |.31| (*SD* = .12).

Table 1. Descriptive Statistics and Intercorrelations for Selected Criterion Variables

Index	Variable	Descriptive Statistics			Intercorrelations			
		Mean	*SD*	Min / Max	1	2	3	4
1	KSA Distance Composite	0.00	0.75	-2.48 / 3.10				
2	Vocational Interests Shape	0.17	0.47	-0.99 / 1.00	-.55			
3	Work Values Shape	0.11	0.54	-1.00 / 1.00	-.42	.36		
4	Work Styles Shape	0.49	0.23	-0.77 / 0.98	-.32	.31	.28	
5	Abs. Med. Salary Diff.	35,578	28,129	0 / 184,260	.28	-.14	-.30	-.17

Notes. Abs. Med. Salary Diff. = Absolute Value of Median Salary Difference. Values were calculated within each O*NET-SOC and then averaged across O*NET-SOCs. For example, intercorrelations were calculated for O*NET-SOC 1 vs. all other O*NET-SOCs, O*NET-SOC 2 vs. all other O*NET-SOCs, etc., and then averaged.

[9] The four O*NET-SOCs missing median salary data were: 27-2011.00 (Actors), 27-2031.00 (Dancers), 27-2042.00 (Musicians and Singers), and 27-2091.00 (Disc Jockeys, Except Radio).

Conceptual Rationale Underlying Choice of Criteria

We used a distance-based metric for the overall KSA composite because absolute levels of knowledge, skill, and ability score (i.e., elevation) differences are meaningful between occupations. Namely, higher levels of knowledge, skill, and ability requirements generally suggest higher job complexity for a given occupation. On the contrary, we did not view overall level differences as conceptually meaningful for vocational interests, work values, or work styles, choosing instead to quantify similarity for these metrics only in terms of relative rank-ordering of profiles. For example, we viewed individuals' ordered preferences on work values (indexed by shape) as important to theoretically distinguish occupations, whereas differentiating occupations that were relatively high in work values versus low in work values (as indexed by elevation) did not hold the same theoretical meaning.

We chose to combine metrics from the cognitive domain (KSAs) into a single criterion composite, whereas we treated the non-cognitive metrics (interests, work values, and work styles) as separate criteria. This decision was motivated by multiple factors. First, individual differences research literature recognizes interests, work values, and work styles (personality) as distinct domains (Hansen & Wiernik, 2018). While knowledges, skills, and abilities are also distinct domains, KSAs as a set can be framed as person-related requirements more clearly stemming from tasks performed on the job, and KSAs represent the traditional core of a job analysis (Harvey, 1991). Second, we expected task-based occupational similarity to differentially relate to the three non-cognitive metrics, given differences between the non-cognitive domain in terms of the composition of their measures or their hypothesized relations with performance. For example, interest assessments often actually include task/work activity-based items (Holland, 1997), theoretical models of job performance suggest personality is more related to contextual rather than task performance (Mototwidlo et al., 1997), and work values measures reflect preferences for "reinforcers" offered by a work environment (Dawis & Lofquist, 1984). Therefore, modeling occupational similarity in vocational interests, work values, and work styles separately allowed us to capture distinct patterns that may emerge.

Evaluation of WB-OSMs Against Occupational Similarity Criteria

To understand how WB-OSMs related to the occupational similarity criteria describe above, we first produced a correlation matrix for each of the 923 data-level occupations that reflected correlations among the WB-OSMs and criteria. To be clear, the unit of analysis for these correlations was occupations. For example, to calculate the correlation between unweighted task similarity and a KSAO similarity for Occupation A, we correlated these two similarity metrics using data from the 922 occupations for which these metrics were calculated (with respect to their similarity to Occupation A). This gave us an estimate of how correlated each pair of similarity metrics were in terms of their rank ordering of occupations for a given target O*NET-SOC (a "target O*NET-SOC" is simply the O*NET-SOC for which a list of related occupations is constructed; each of the 923 data-level occupations is the target O*NET-SOC for its respective list of related occupations). Once these correlation matrices were produced for each of the 923 occupations, we summarized the correlations for each pair of metrics across occupations (e.g., mean correlation, standard deviation of correlations, and range of correlations). To the extent that work performed in occupations drives factors such as salary and importance of KSAs, interests, work value, and work styles, we expected to see strong relations between the WB-OSMs and criteria reflecting similarity on the factors above. For a summary of mean descriptive statistics and mean intercorrelations for our set of criteria and WB-OSMs, see Appendix A.

Prior to examining the results, we expected that similarity of task statements across occupations, relative to similarity of DWAs or occupation descriptions, would provide the most robust basis for estimating the similarity of occupations in terms of work performed. As such, we expected the correlations between task-based similarity indices and the criteria to be stronger than both (a) the correlations between DWA-based similarity indices and the criteria and (b) correlations between occupation description-based similarity indices and the criteria.

Indeed, one question we aimed to answer was whether the DWA-based similarity or occupation-description based similarity metrics have incremental value above task-based similarity metrics for predicting the criteria. Additionally, we also aimed to address whether weighted versions of the task- and DWA-based similarity metrics offer any incremental value over their unweighted versions for predicting the criteria. Before addressing these questions, however, we used our correlational results to select the best method for quantitatively representing the text.

Selection of a Final Method for Quantitatively Representing Text

We reviewed the correlations between the WB-OSMs and criteria to identify the method for quantitatively representing text that performed best. The TF-based cosines were the least predictive of the five criteria, followed by the TF-IDF-based cosines; this was expected, as these methods only reflect shared usage of words across documents and cannot account for the semantic similarity of different words. The GloVe (unweighted), GloVe (TF-IDF weighted), USE-DAN, and SBERT methods all avoided this limitation and produced larger correlations with the five criteria. These four semantically sensitive methods had comparable magnitudes of correlations with the five criteria; the two GloVe-based methods were slightly better than the other methods at predicting the KSA distance composite, but performance across the other four criteria was quite similar.

Given how similarly each of the semantically sensitive methods correlated with our criteria, we based our choice of method primarily on differences in the sophistication of their language model and their ease of use. In this case, both of those features were positively correlated: SBERT and USE-DAN were the easiest-to-use methods, and they also used the most nuanced language models, as both methods account for the contexts in which words are used and can interpret novel words. The GloVe methods, on the other hand, are much more complicated to implement, cannot account for word context, and can only be used with a fixed (albeit large) set of words. We chose to use SBERT as our final method for computing similarity metrics, as it performed similarly to USE-DAN for task statements and DWAs in terms of convergence with criteria, but it performed better than USE-DAN for occupation descriptions.

Incremental Validity Analyses for Predicting Criteria

With the final method for quantitatively representing the text in hand, we next conducted a series of incremental validity analyses for each target O*NET-SOC (again using occupations as the unit of analysis) to address the questions above more directly. Rather than conduct these analyses by fitting hierarchical regression models (e.g., including predictor A in the first step of a model and predictors A and B in the second step, to examine the increment of B over A), we evaluated incremental validity by comparing criterion-related validity of composites computed as equally weighted averages of the WB-OSM variables. We adopted this strategy because it would not be practical to recommend that the Center use a different regression weighted composite of task-, DWA-, and/or description-based similarity metrics for each occupation. Thus, our strategy here reflected realistic ways these composites might be formed for operational use.

We used SBERT-based WB-OSMs to evaluate the following set of incremental validity contrasts for the 4,615 combinations of data-level O*NET-SOCs and criteria:

- Contrast 1: Examine change in R^2 from adding weighted task similarity to unweighted task similarity to determine if weighting adds value.
- Contrast 2: Examine change in R^2 from adding weighted DWA similarity to unweighted DWA similarity to determine if weighting adds value.
- Contrast 3a: Examine change in R^2 from adding DWA similarity to task similarity to determine if DWAs add value over tasks.
- Contrast 3b: Examine change in R^2 from adding occupation-description similarity to task similarity to determine if occupation descriptions add value over tasks.
- Contrast 4: Examine change in R^2 from adding occupation-description similarity to task and DWA similarity to determine if occupation descriptions add value over tasks and DWAs.

After computing these contrasts, we examined patterns in the distributions of differences in R^2 values to determine which predictors added value (see Table 2). These incremental validity analyses revealed that weighted similarity metrics did not explain additional variance in our criteria beyond the unweighted similarity metrics (Contrasts 1 and 2); this lack of incremental validity is also clearly supported by the unity or near-unity correlations between unweighted and weighted similarity metrics shown in Appendix A. Based on Contrasts 1 and 2, we only used unweighted metrics when comparing the incremental predictive value of DWAs and occupation descriptions (Contrasts 3a, 3b, and 4). We found that adding DWA similarity to task similarity provided modest increments in prediction; however, combining occupation description-based similarity with the task and DWA similarity metrics did not reliably improve prediction. These results supported using a WB-OSM composite computed as an average of task- and DWA-based WB-OSMs.

Table 2. Distributions of Differences in R^2 Values for Model Contrasts Using SBERT Cosines

Criterion	*k*	Contrast	*M*	*SD*	*Min*	5th %ile	*Mdn*	95th %ile	*Max*
KSA Distance Composite	873	1	0.00	0.00	-0.01	0.00	0.00	0.00	0.01
		2	0.00	0.02	-0.13	-0.04	0.00	0.03	0.07
		3a	0.03	0.05	-0.21	-0.06	0.03	0.12	0.22
		3b	-0.03	0.05	-0.27	-0.12	-0.03	0.06	0.16
		4	-0.02	0.04	-0.22	-0.07	-0.02	0.04	0.12
Vocational Interests Shape	874	1	0.00	0.00	-0.01	0.00	0.00	0.00	0.03
		2	0.00	0.02	-0.13	-0.03	0.00	0.03	0.10
		3a	0.04	0.05	-0.22	-0.03	0.04	0.13	0.27
		3b	0.00	0.06	-0.27	-0.09	0.00	0.10	0.16
		4	0.00	0.04	-0.16	-0.07	0.00	0.06	0.11

Table 2. (Continued)

Criterion	*k*	Contrast	*M*	*SD*	*Min*	5th %ile	*Mdn*	95th %ile	*Max*
Work Values Shape	874	1	0.00	0.00	-0.01	0.00	0.00	0.00	0.01
		2	-0.01	0.01	-0.08	-0.03	0.00	0.01	0.04
		3a	0.02	0.03	-0.11	-0.03	0.01	0.07	0.16
		3b	-0.01	0.03	-0.13	-0.06	-0.01	0.04	0.11
		4	-0.01	0.02	-0.09	-0.05	-0.01	0.02	0.08
Work Styles Shape	873	1	0.00	0.00	0.00	0.00	0.00	0.00	0.01
		2	0.00	0.01	-0.03	-0.01	0.00	0.02	0.10
		3a	0.01	0.03	-0.09	-0.02	0.01	0.06	0.13
		3b	0.00	0.02	-0.10	-0.04	0.00	0.04	0.10
		4	0.00	0.02	-0.11	-0.03	0.00	0.03	0.06
Abs. Med. Salary Diff.	919	1	0.00	0.00	-0.01	0.00	0.00	0.00	0.00
		2	0.00	0.01	-0.05	-0.02	0.00	0.01	0.03
		3a	0.02	0.02	-0.08	-0.01	0.01	0.05	0.13
		3b	0.00	0.02	-0.11	-0.04	0.00	0.02	0.08
		4	-0.01	0.02	-0.09	-0.04	0.00	0.01	0.06

Note. *k* represents the number of data-level O*NET-SOCs that had data available for the focal criterion. Contrasts 3a, 3b, and 4 are only reported for unweighted cosines because Contrasts 1 and 2 did not support that use of weighting added value.

Calculating an Initial Occupational Relatedness Composite

Based on the results above, we decided to calculate an initial occupational relatedness composite based only on task- and DWA-based similarity. Specifically, we computed our composite as an unweighted average of task-based similarity and DWA-based similarity.

Phase 2: Review of the Initial Related Occupations Lists and Adjustments to the Initial Occupational Relatedness Composite

Using the initial occupational relatedness composite described above, we identified the top-25 most similar potentially related occupations for each of the 923 data-level occupations included within the O*NET-SOC 2019 taxonomy. Two HumRRO researchers and a senior member of the Center's staff reviewed these preliminary results to determine the viability of the task- and DWA-similarity composite. We found that, although this composite was able to produce sensible rank-orderings of related occupations for many target O*NET-SOCs, it did not function well in subsets of occupations; specifically, it performed poorly when occupations included similarly worded tasks and DWAs but had important differences in their context or subject matter, suggesting that the composite was deficient in some way. We found clear examples of this potential deficiency in teaching and managerial occupations.

For teaching occupations, the initial metric tended to recommend other types of teaching occupations, even if those occupations involved entirely different academic disciplines. Most postsecondary teaching occupations included the same set of occupations at the top of their lists, such as:

- Tutors (25-3041.00)
- Career/Technical Education Teachers, Postsecondary (25-1194.00)
- Adult Basic Education, Adult Secondary Education, and English as a Second Language Instructors (25-3011.00)
- Self-Enrichment Teachers (25-3021.00)

These four occupations contain some of the most generic teaching-focused task statements and DWAs and were therefore the easiest to match with other teaching-focused occupations. However, we were expecting to see more related occupations outside of teaching that were in the same disciplines as the target O*NET-SOCs. For example, we anticipated that Biological Science Teachers, Postsecondary (25-1042.00) would be related to Biologists (19-1029.04); however, most of the recommendations were teaching occupations in other disciplines, while Biologists were ranked 163rd in the complete list of 922 potentially related occupations.

For managerial occupations, the initial metric tended to recommend other types of management occupations, regardless of the field in which the managers work. This is a less extreme version of the problem we noticed for teaching occupations, as people-managing skills arguably transfer more easily across fields/disciplines than do teaching skills. A teacher would need to receive extensive training in a new field to teach it, but a manager may be able to supervise technical workers in a given field without sharing their background. However, many managers obtain their jobs by first working in a variety of technical roles, and we expected that our lists of related occupations would reflect this technical background (e.g., Financial Managers [11-3031.00] should be related to other occupations within the finance realm).

Due to this apparent deficiency of our initial WB-OSM composite, we explored ways to augment our composite and incorporate other types of indicators about how occupations are related.

Refining the Initial Occupational Relatedness Composite

Based on the observations above, we took several steps to refine the initial relatedness composite in ways that would address issues revealed through the initial review. In this section, we review the refinements we considered, and the results of our evaluation of those refinements.

Augmentation of the Relatedness Composite

When we recognized the deficiency in the initial relatedness composite, we took stock of other data available in the O*NET database and considered ways we could use it to compute metrics we could add to our composite. Our focus was on identifying as few metrics as necessary to supplement the task- and DWA-based WB-OSM composite, and we decided that metrics summarizing similarity in knowledges and alternate titles would be best suited to address the types of deficiencies observed above. The rationale behind our focus on knowledges and alternate titles is they both help differentiate occupations in terms of discipline-specific subject matter, and it is in this area where simply relying on only tasks and DWAs was problematic for certain types of occupations.

To quantify knowledge-based similarity, we computed cosines between profiles of importance ratings across O*NET's 33 knowledge domains for each pair of O*NET-SOCs. The task of qualifying similarity in alternate titles was more involved. To evaluate similarity in alternate titles, we used the following NLP-based approach to compute cosines:

1. For each O*NET-SOC's list of alternate titles, we extracted the unique words that appeared in the titles and counted the number of times each word occurred in the list.
2. We queried the database of pre-trained GloVe embeddings to find numeric representations of each unique word.
3. We computed TF-IDF weights for the words that appeared in alternate title lists, and we conducted this process separately within each job family. We computed weights within job families because it allowed us to down-weight words that are common in certain contexts but less common in others (e.g., "teacher" is very common within the "Educational Instruction and Library" job family, but uncommon in other job families). By differentially weighting words within job families, we aimed to increase our ability to detect similarity between occupations that share a discipline-specific subject matter but are in different job families (e.g., Biological Science Teachers, Postsecondary are in the "Educational Instruction and Library" job family while Biologists are in the "Life, Physical, and Social Science" job family, and down-weighting common teacher words could help us to find their biology-based relatedness).
4. We used the GloVe embeddings and TF-IDF weights compute a single weighted-average vector of embeddings per O*NET-SOC.
5. Finally, we computed a cosine between the embeddings from each pair of O*NET-SOCs.

We computed our revised relatedness composite as an unweighted average of three variables:

- *Z* scores for the initial composite of task- and DWA-based similarity,
- *Z* scores of cosines between knowledge profiles, and
- *Z* scores of cosines between embeddings representing alternate titles.

For all metrics, we standardized the values across all pairs of related occupations that had data for the metric. After combining these three standardized variables into a composite, we standardized the composite across all 851,006 occupation pairs to give it a more interpretable scaling. In symbolic terms, we computed this composite using the following formula:

$$Relatedness = Z\left(\frac{Z[WBOSM_{Tasks\ \&\ DWAs}] +\ Z\left[Cosine_{Knowledges}\right] + Z[Cosine_{Alternate\ Titles}]}{3^{*}}\right)$$

*If knowledge importance ratings were unavailable for at least one O*NET-SOC involved in a given pair, this 3 was replaced with a 2 to ensure the available values were averaged correctly. Knowledge ratings were missing for 50 data-level O*NET-SOCs, which produced a total of 89,750 missing cosines across occupation pairs.

This new composite helps to make up for the deficiencies of our initial metric by accounting for three important contributors to occupational similarity: what people in the occupations *do*, what they *know*, and what they are *called*. In addition to making up for deficiencies in the initial relatedness composite, this three-part formulation of relatedness quantifies occupational relatedness using features of occupations that are intuitively relevant to job similarity.

Evaluation of the Reformulated Relatedness Composite

After augmenting our occupational relatedness metric to account for similarity between occupations' knowledges and titles, we conducted two types of evaluations: (1) an informal preliminary review by two HumRRO researchers to determine whether the updated rank-ordering of related occupations made sense across a variety of O*NET-SOCs and (2) a more structured review by the Center's staff, featuring the related-occupations lists for a diverse sample of O*NET-SOCs.

In the informal preliminary review, we observed that the updated composite appeared to improve the quality of related occupations for sets of O*NET-SOCs that had seemed problematic when we reviewed the performance of the initial relatedness metric. For example, teaching occupations emerged as more similar to other occupations within their respective disciplines, and managerial occupations emerged as more similar to other occupations within their respective fields.

Based on our informal review, we proceeded to conduct a more structured review involving member of the Center's staff. We sampled a total of 60 of the 923 data-level occupations from the O*NET-SOC 2019 taxonomy to include in this review. Of the 60 O*NET-SOCs, 50 were selected through random sampling (stratified by job family and job zone[10] to ensure a diverse and representative set of occupations), five were selected because the average relatedness score of their 10 most similar related occupations was very low, and five were selected because the average relatedness score of their 10 most similar related occupations was very high. This sampling strategy allowed us to examine both a random and strategic subset of occupations.

For each of the 60 sampled O*NET-SOCs, we constructed an Excel-based review sheet that listed (1) the 10 highest-similarity related occupations, (2) occupations that were related to the target O*NET-SOCs through the structure of the O*NET-SOC 2019 taxonomy (i.e., O*NET-SOCs that were considered "parent-child" pairs or "sibling" occupations), and (3) the 11th through 25th ranked related occupations. We called the O*NET-SOCs from the first category "primary" related occupations and we called those from the second and third categories "alternate" related occupations. Center staff reviewed each set of 10 primary related occupations and determined if any were not a good fit for the target occupations; if so, they selected one of the alternate related occupations to use in place of the poorly fitting primary related occupation.

Out of the related occupation lists for the 60 sample O*NET-SOCs, the Center's staff members recommended replacing only 35 primary related occupations (5.8% of the 600 total primary related occupations). These replacements were distributed across 23 of the 60 O*NET-SOCs; the Center's staff did not recommend replacing any primary related occupations for 37 of the 60 O*NET-SOCs. It is interesting to note that reviewers recommended replacing the same number of related occupations from the target O*NET-SOCs we sampled due to their having low and high average relatedness scores; for each of these strategically sampled sets of O*NET-SOCs, the reviewers recommended three total replacements. This provided encouraging evidence that the rank ordering of relatedness scores within the list of related occupations for a given O*NET-SOC was more important than the magnitudes of the scores within those lists. In other words,

[10] Job zone categories included within O*NET identify groups of occupations that are similar with respect to the education, experience, and on-the-job training needed to do the work (Rivkin & Craven, 2021). Job zones are coded using values ranging from 1 (occupations that need little or no preparation) to 5 (occupations that need extensive preparation).

even O*NET-SOCs that have low magnitudes of similarity with other occupations can still have intuitive and meaningful rank-ordered lists of related occupations.

Narrative feedback from the Center indicated that the sets of primary related occupations were generally very good and that replacements were often made based on differences in job zones between the target O*NET-SOCs and their primary related occupation. Based on this review, the Center recommended that HumRRO continue using the revised relatedness composite to create lists of related occupations, and we proceed with processing data from the most current available database at the time (the O*NET 26.1 Database; National Center for O*NET Development, 2021b) to develop new related occupations for inclusion in the upcoming O*NET 26.3 Database release.

Phase 3: Construction of an Occupational Relatedness Threshold

Prior to processing the O*NET 26.1 Database (National Center for O*NET Development, 2021b), we conducted a final piece of research using the O*NET 25.3 Database (National Center for O*NET Development, 2021a). We designed a review activity intended to identify a threshold value that could be informative for differentiating occupation pairs that have a good chance of being truly related from those that have a lower chance. We anticipated that such a threshold value could be a useful component of evaluative analyses after we constructed the final related occupations matrix.

We took the complete set of unique O*NET-SOC pairs (*N* = 425,503) from the O*NET 25.3 Database (National Center for O*NET Development, 2021a), sorted them by relatedness score from highest to lowest, converted the relatedness scores to percentile ranks, and randomly sampled 20 occupation pairs from each percentile to include in a review activity. We determined that 20 pairs per percentile would provide an adequate representation of the full set of related occupations while keeping the activity manageable in scope. To further constrain the scope of the activity and prevent the list of occupation pairs from intimidating reviewers, we set a conservative lower bound on the list of occupation pairs: We eliminated pairs below the 60th percentile because it was clear those pairs were generally too dissimilar to be considered related by an informed reviewer. This left a total of 800 occupation pairs for inclusion in the review process.

We recruited five HumRRO researchers with PhDs in industrial and organizational psychology to serve as reviewers. We gave them the list of 800 occupation pairs to and asked each of them to make binary judgments about whether each pair was related, with instructions to stop recording judgments when they reached a point in the list where they no longer observed acceptable levels of relatedness. This exercise was a variation of “bookmarking” (i.e., an activity where reviewers draw a line or “place a bookmark” at the point in a list where the list items are no longer suitable for a given purpose) in which the reviewers both make an overall decision to stop their review while making binary judgments along the way.

The reviewers ended the exercise between the 89th and 95th percentiles of occupation pairs, their median stopping point was the 93rd percentile, and their mean stopping point was the 92nd percentile. We determined that the median stopping point should be the key determinant of the relatedness cut off and we set the threshold at 1.61649, which represented the bottom of the 93rd percentile in our analyses. We planned to use this threshold as a key part of our evaluation step after building our final related occupations matrix.

Phase 4: Construction and Review of the Final Related Occupations Matrix

We used the final occupational relatedness composite to identify the top-25 related occupations for each of the 923 data-level occupations included in the O*NET 26.1 Database (National Center for O*NET Development, 2021b). As with the previous review activity we conducted with the Center, we constructed review files that included the top-25 related occupations for each target O*NET-SOC, as well as any occupations that were related to the target O*NET-SOC via parent-child or sibling relations within the O*NET-SOC 2019 taxonomy.

We provided these review file to the Center for evaluation and comments. Specifically, the Center's staff provided the following input for each target O*NET-SOC's set of related occupations:

- Should any of the top-10 related occupations be replaced with an alternate related occupation?
- If a top-10 related occupation needed to be replaced, should it be retained as a supplemental related occupation, or dropped from lists of related occupations displayed on O*NET OnLine (and potentially the My Next Move sites)?

Of the 9,230 total related occupations that were in the top-10 lists across the 923 data-level O*NET-SOCs, reviewers recommended replacing 1,272 related occupations with alternate occupations (13.78%); these replacements were distributed across 579 of the 923 O*NET-SOCs examined. This rate of replacements was more than double the 5.8% observed in the earlier review of 60 sampled O*NET-SOCs, but it is important to note that the reviewers who participated in this exercise were not involved in the development of the relatedness composite, nor the earlier review exercise.

Table 3 shows cumulative frequencies of replacements across the 923 data-level O*NET-SOCs; 344 (37.27%) had no replacements and 575 (78.77%) had two or fewer replacements. Despite the higher-than-expected rate of replacements, reviewers only recommended dropping five (0.39%) of the 1,272 replaced related occupations from the lists of related occupations that would potentially be displayed on O*NET's websites; please see Table 4 for a listing of these five dropped related occupations. After accounting for the five dropped related occupations, the remaining 1,267 (99.61%) were simply demoted from "primary" related occupations (i.e., occupations that will be displayed by default in online resources) to "supplemental" related occupations (i.e., occupations that will not initially be displayed in online resources but can be viewed if users click an option to show additional occupations).

Table 3. Cumulative Frequencies of Empirical Top-10 Related Occupations Replaced by Reviewers

Frequency of Replacements Per Target O*NET-SOC	Non-Cumulative Summary of Target O*NET-SOCs		Cumulative Summary of Target O*NET-SOCs with a Number of Replacements Less Than or Equal to Row Frequency		Cumulative Summary of Target O*NET-SOCs with a Number of Replacements Greater Than or Equal to Row Frequency	
	k	%	*k*	%	*k*	%
0	344	37.27	344	37.27	923	100.00
1	231	25.03	575	62.30	579	62.73
2	152	16.47	727	78.77	348	37.70
3	97	10.51	824	89.28	196	21.23
4	64	6.93	888	96.21	99	10.72
5	22	2.38	910	98.59	35	3.79
6	11	1.19	921	99.78	13	1.41
7	2	0.22	923	100.00	2	0.22

Table 4. Summary of Empirical Top-10 Related Occupations Dropped by Reviewers

O*NET-SOC Code of Target Occupation	Title of Target Occupation	O*NET-SOC Code of Dropped Related Occupation	Title of Dropped Related Occupation
49-3022.00	Automotive Glass Installers and Repairers	51-9083.00	Ophthalmic Laboratory Technicians
19-4051.02	Nuclear Monitoring Technicians	29-2036.00	Medical Dosimetrists
39-5093.00	Shampooers	29-1213.00	Dermatologists
39-5093.00	Shampooers	29-2055.00	Surgical Technologists
51-9151.00	Photographic Process Workers and Processing Machine Operators	51-9083.00	Ophthalmic Laboratory Technicians

After the Center completed their review and we accounted for all their recommended alterations to the related occupations lists, we constructed final work products for inclusion in the O*NET 26.3 Database (see Appendix C for more information about these files' contents):

- an "Operational Related Occupations Matrix" that can be used to identify which related O*NET-SOCs to display on O*NET's suite of web sites that includes 10 primary related occupations and 10 supplemental primary related occupations for each O*NET-SOC, and
- a "Related Occupations Research Dataset" that contains quantitative similarity information about how each of the 923 data-level O*NET-SOCs relates to each of the other 922 data-level O*NET-SOCs.

Phase 5: Evaluation of the Final Related Occupations Matrix

Once we constructed the final set of operational related occupations, we conducted follow-up analyses to evaluate the matrix with respect to (a) the threshold similarity value we established in our bookmarking activity and (b) how primary related occupations differed from other occupations on external criteria.

Comparisons to the Bookmark Similarity Threshold

We began our evaluation by comparing the relatedness scores of occupation pairs in the new matrix to the threshold value from our bookmarking exercise. As a baseline for understanding the level of similarity between related occupations in the new matrix, we also applied our analyses to the related occupations listed in O*NET's "Career Starters" and "Career Changers" related occupations matrices from the O*NET 26.1 Database (National Center for O*NET Development, 2021b). Given our complete redesign of the process for identifying related occupations, we did not expect substantial alignment between the evaluation results for the new operational related occupations matrix and the Career Starters or Career Changers matrices; our goal in making these comparisons was to understand how the new matrix compares to those that preceded it with regard to the new relatedness composite.

Table 5 summarizes differences in the numbers of related occupations across target O*NET-SOCs across the old and new matrices. In the new operational matrix (and the empirical top-10 occupations that we sent to the Center for review), every data-level O*NET-SOC has exactly 10 "primary" related occupations. In the Career Starters and Career Changers, however, target O*NET-SOCs could be related to fewer occupations—as few as eight in the Career Starters matrix and as few as one in the Career Changers matrix.[11]

Table 5. Frequencies of Target O*NET-SOCs by Number of Related Occupations and Related Occupations Matrix

Number of Related Occupations	Frequency of Target O*NET-SOCs by Related Occupations Matrix			
	Career Starters	Career Changers	Empirical Top-10	Final New Matrix
1	0	1	0	0
2	0	5	0	0
3	0	3	0	0
5	0	5	0	0
6	0	10	0	0
7	0	26	0	0
8	26	74	0	0
9	180	188	0	0
10	571	465	923	923
Total	777	777	923	923

[11] Updates to the occupational classification structure after the original development work led to instances where previously related occupations were dropped or significantly modified and therefore not included.

Table 6 shows descriptive statistics for the mean relatedness scores of the related occupations across target O*NET-SOCs from each matrix. Table 6 also includes summaries for versions of the empirical top-10 related occupations and the final new matrix from this research that are limited to the same 777 O*NET-SOCs included in the in the Career Starters and Career Changers matrices; these limited versions of the matrices built using our new methodology support more direct comparisons with the old matrices. The empirical top-10 matrix and the final new matrix were informed by the new relatedness composite, so they will naturally tend to have higher mean relatedness scores than the old matrices; however, comparing the means across matrices can help to understand how much the new composite differs from the way occupational similarity was quantified in past research.

The grand mean of relatedness scores for the empirical top-10 related occupations (limited to the O*NET-SOCs included in the old matrices) was much higher than the grand means for both the Career Starters (d = 1.73) and Career Changers (d = 1.02) matrices. After the Center reviewed the empirical top-10 related occupations and made their substitutions, the grand mean for primary related occupations dropped by only .09 standard deviations. Given this small change, the grand mean of relatedness scores for the final matrix (again limited to the O*NET-SOCs included in the old matrices) was also much higher than the grand means for both the Career Starters (d = 1.63) and Career Changers (d = 0.93) matrices. These patterns of mean differences suggest that our relatedness composite was more similar to the metrics that informed the Career Changers matrix than those that informed the Career Starters matrix, but the differences still outweigh their similarities.

Table 6. Summary of Mean Relatedness Scores for Related O*NET-SOCs by Related Occupations Matrix

Matrix	*k*	*M*	*SD*	*Min*	5th *%ile*	*Mdn*	95th *%ile*	*Max*	Means Above Threshold	
									k	%
Career Starters	777	1.56	0.58	-0.21	0.52	1.60	2.42	3.09	432	55.60
Career Changers	777	1.94	0.58	-0.26	0.90	1.98	2.85	3.34	596	76.71
Empirical Top-10	923	2.51	0.50	0.70	1.66	2.52	3.31	3.99	897	97.18
Final New Matrix	923	2.47	0.51	0.70	1.59	2.46	3.26	3.97	888	96.21
Empirical Top-10 (Limited)	777	2.49	0.49	0.78	1.67	2.47	3.25	3.67	758	97.55
Final New Matrix (Limited)	777	2.45	0.50	0.77	1.63	2.43	3.23	3.67	750	96.53

Note. The descriptive statistics in this table represent distributions of mean relatedness scores for the target O*NET-SOCs tallied in Table 5. Matrices identified as “limited” were reduced to include only the 777 O*NET-SOCs that were represented in O*NET’s Career Starters and Career Changers matrices from the O*NET 26.1 Database (National Center for O*NET Development, 2021b). The results for “Means Above Threshold” represent the number and percentage of target O*NET-SOCs whose mean relatedness scores exceeded the threshold of 1.61649 established in our bookmarking activity.

Table 7 summarizes the proportions of related occupations that exceeded our bookmark threshold across target O*NET-SOCs from each matrix. Consistent with the large mean differences in relatedness scores between the old and new matrices described above, there were large differences in the rates at which target O*NET-SOCs’ related occupations cleared our threshold value. Compared to the rates for the empirical top-10 related occupations (mean of 94% of related occupations above the threshold) and the final new matrix (mean of 93% of related occupations above the threshold; 92% after limiting the target O*NET-SOCs to those included in the Career Starter and Career Changer matrices), rates of related occupations were much lower for the Career Starters matrix (49%) and the Career Changers matrix (67%).

Table 7. Summary of Proportions of Related O*NET-SOCs that Exceeded the Bookmarking Threshold by Related Occupations Matrix

Matrix	*k*	*M*	*SD*	*Min*	5th %ile	*Mdn*	95th %ile	*Max*	Sets of Related Occupations Completely Above Threshold *k*	Sets of Related Occupations Completely Above Threshold %
Career Starters	777	0.49	0.30	0.00	0.00	0.50	0.90	1.00	38	4.89
Career Changers	777	0.67	0.30	0.00	0.10	0.70	1.00	1.00	207	26.64
Empirical Top-10	923	0.94	0.18	0.00	0.50	1.00	1.00	1.00	818	88.62
Final New Matrix	923	0.93	0.19	0.00	0.40	1.00	1.00	1.00	753	81.58
Empirical Top-10 (Limited)	777	0.94	0.18	0.00	0.50	1.00	1.00	1.00	684	88.03
Final New Matrix (Limited)	777	0.92	0.19	0.00	0.40	1.00	1.00	1.00	628	80.82

Note. The descriptive statistics in this table represent distributions of proportions of related occupations that met or exceeded the threshold of 1.61649 established in our bookmarking activity for the target O*NET-SOCs tallied in Table 5. Matrices identified as "limited" were reduced to include only the 777 O*NET-SOCs that were represented in O*NET's Career Starters and Career Changers matrices from the O*NET 26.1 Database (National Center for O*NET Development, 2021b). The results for "Sets of Related Occupations Completely Above Threshold" represent the number and percentage of target O*NET-SOCs for which all related occupations exceeded the threshold value.

Overall, the relations among O*NET-SOCs featured in the new operational related occupations matrix compare favorably to the threshold we established in our bookmarking activity. On average, across the 923 data-level O*NET-SOCs, 93% of the primary related occupations had relatedness scores that exceeded the threshold. Additionally, 96.21% of target O*NET-SOCs had mean relatedness scores for their primary related occupations that exceeded the threshold value, and 81.58% of target O*NET-SOCs had sets of primary related occupations in which all 10 of the primary relations exceeded the threshold value. As shown the cumulative frequency distribution in Table 8, only 5.85% of target O*NET-SOCs in the final operational matrix had sets of primary related occupations in which fewer than five of the primary relations had relatedness scores above the threshold value. Cumulative frequency distributions for the rest of the matrices we evaluated earlier in this section are available in Appendix B.

*Table 8. Cumulative Frequency Distribution for Proportions of Related O*NET-SOCs that Exceeded the Bookmarking Threshold in the Final Operational Related Occupations Matrix*

Proportion of Related Occupations Above Threshold	Non-Cumulative Summary of Target O*NET-SOCs		Cumulative Summary of Target O*NET-SOCs With Proportions Less Than or Equal to Row Proportion		Cumulative Summary of Target O*NET-SOCs With Proportions Greater Than or Equal to Row Proportion	
	k	%	*k*	%	*k*	%
0.0	4	0.43	4	0.43	923	100.00
0.1	3	0.33	7	0.76	919	99.57
0.2	16	1.73	23	2.49	916	99.24
0.3	17	1.84	40	4.33	900	97.51
0.4	14	1.52	54	5.85	883	95.67
0.5	16	1.73	70	7.58	869	94.15
0.6	16	1.73	86	9.32	853	92.42
0.7	17	1.84	103	11.16	837	90.68
0.8	21	2.28	124	13.43	820	88.84
0.9	46	4.98	170	18.42	799	86.57
1.0	753	81.58	923	100.00	753	81.58

Differences Between Primary Related Occupations and Other Occupations on External Criteria

The first step of our evaluation produced favorable results and demonstrated that the primary related occupations in the new matrix exceeded our bookmark-based threshold at a very high rate. Next, we evaluated the magnitudes by which each target O*NET-SOC's 10 primary related occupations differed from the other 912 occupations with respect to external criteria. For these comparisons, we defined the "primary related occupations" as those that are ranked among the top-10 related occupations after expert review and we defined the "non-primary occupations" as the other 912 occupations. Specifically, we compared primary related occupations and non-primary occupations in terms of their similarity to the target O*NET-SOCs' skill, ability, interest, work style and work value profiles, as well as their salaries and job zones.[12] These comparisons were based on our earlier criterion analyses and included:

- Skill Importance Profile Correlation
- Ability Importance Profile Correlation
- Interest Profile Correlation
- Work Style Profile Correlation
- Work Value Profile Correlation

[12] We did not evaluate knowledge profile similarity, as knowledge similarity factored into our process for calculating the relatedness scores used to identify related occupations.

- Median Salary Absolute Difference
- Job Zone Absolute Difference

Additionally, to understand the upper limits of the effects we could expect for these external criteria, we compared primary related occupations and non-primary occupations based on their relatedness scores. Given that the relatedness scores are the most influential factor in determining which occupations are categorized as primary related occupations, primary and non-primary occupations should exhibit very large differences on this variable. We used two methods to compare primary related occupations to other occupations: Mean difference analyses and variance reduction analyses.

Mean Difference Analyses

Our process for conducting mean difference analyses included three main steps. In Step 1, we identified the primary related occupations and non-primary occupations for each target O*NET-SOC and computed the means and standard deviations of the criterion variables for each of these sets of occupations. The result of this was up to 923 sets of means and *SD*s for primary related occupations and up to 923 sets of means and *SD*s for non-primary occupations. In Step 2, we used the means and *SD*s from Step 1 to compute standardized mean differences (i.e., *d* values) between the primary related occupations and the non-primary occupations of occupations for each target O*NET-SOC. The result of this was up to 923 *d* values per criterion variable. In Step 3, the final step, we (a) aggregated the means and *SD*s from Step 1 to describe the distributions of criterion values for primary related occupations and non-primary occupations and (b) summarized the *d* values from Step 2 to characterize the differences between primary related occupations and non-primary occupations.

Table 9 shows the aggerated descriptive statistics for primary related occupations and non-primary occupations for relatedness scores and each of the external criteria. Note that some O*NET-SOCs did not have data on the external criteria, so the number of occupations included in our analyses is less than 923 in some cases. As there are two levels of analysis involved in our evaluation—within target O*NET-SOCs and between target O*NET-SOCs—Table 9 presents information relevant to both levels. Due to our analyses involving two levels of analysis, there are several different ways one might calculate summary statistics such as means and standard deviations. Thus, prior to discussing the results in Table 9, we offer more details about how the summary statistics in that table were calculated.

Recall from Step 1 of our three-step analysis process that we computed descriptive statistics separately for primary related occupations and non-primary occupations. For primary related occupations and non-primary occupations, the grand mean (M_{Grand}) represents the mean of all within-occupation criterion score means across target O*NET-SOCs, while the *SD* of means (SD_{Means}) represents the standard deviation of those same within-occupation criterion score means. Thus, the grand means and *SD*s of means summarize distributions of criterion scores aggregated at the level of target O*NET-SOCs (as opposed to individual pairings of target O*NET-SOCs and related occupations). The pooled *SD* (SD_{Pooled}), on the other hand, represents the square root of the average variance (i.e., SD^2) in criterion scores across target O*NET-SOCs. Thus, while the *SD*s of means characterize the *between*-occupation variability of criterion scores among target O*NET-SOCs, the pooled *SD*s characterize the average *within*-occupation variability of criterion scores across target O*NET-SOCs.

*Table 9. Criterion Descriptive Statistics for Primary and Non-Primary Related Occupations Across Data-Level O*NET-SOCs*

Criterion	*k*	Primary Related Occupations				Non-Primary Occupations			
		Mean *n*	M_{Grand}	SD_{Means}	SD_{Pooled}	Mean *n*	M_{Grand}	SD_{Means}	SD_{Pooled}
Relatedness Score*	923	10.00	2.47	0.51	0.29	912.00	-0.03	0.34	0.91
Skill Importance Profile Correlation	873	8.79	0.87	0.07	0.07	815.97	0.62	0.12	0.23
Ability Importance Profile Correlation	873	8.79	0.89	0.07	0.06	815.97	0.68	0.11	0.21
Interest Profile Correlation	874	8.80	0.79	0.16	0.18	817.85	0.16	0.20	0.47
Work Style Profile Correlation	873	8.79	0.71	0.12	0.13	815.97	0.49	0.11	0.23
Work Value Profile Correlation	874	8.80	0.60	0.25	0.29	817.85	0.11	0.14	0.55
Median Salary Absolute Difference	923	10.00	17,229.30	14,910.03	16,529.43	912.00	36,555.69	20,986.18	29,187.82
Job Zone Absolute Difference	923	10.00	0.39	0.31	0.46	912.00	1.30	0.33	0.95

Note. *k* = Number of target data-level O*NET-SOCs with data available for the criterion. Mean *n* = Mean number of primary related O*NET-SOCs with data available for the criterion. M_{Grand} = Mean of within-occupation criterion score means across target O*NET-SOCs. SD_{Means} = Standard deviation of within-occupation mean criterion scores across target O*NET-SOCs. SD_{Pooled} = Square root of the average within-occupation variance in criterion scores across target O*NET-SOCs.

* Relatedness scores are not external criteria; they are included here as a point of comparison because they are the variable most directly related to the distinction between primary and non-primary related occupations.

As a concrete example of what these statistics represent, consider the results for the "Skill Importance Profile Correlation" criterion in Table 9. There were 873 target O*NET-SOCs for which criterion scores were available and, across those 873 target O*NET-SOCs, criterion scores were available for an average of 8.79 primary related occupations and an average of 815.97 non-primary occupations. The M_{Grand} and SD_{Means} values for primary related occupations indicate that, after computing the mean criterion score for primary related occupations within each target O*NET-SOC, the mean of these means was 0.87 and the *SD* of these means was 0.07. Likewise, the mean of the means for non-primary occupations within each target O*NET-SOC was 0.62 and the *SD* of these means was 0.12. The SD_{Pooled} value for primary related occupations indicates that, after computing the variance of criterion scores for primary related occupations within each target O*NET-SOC, the square root of the mean of these variances was 0.07. Likewise, the square root of the mean of the variances for non-primary occupations within each target O*NET-SOC was 0.23.

Table 10 summarizes the standardized mean differences between primary related occupations and non-primary occupations across the target O*NET-SOCs, and Figure 1 depicts the distributions of *d* values graphically. Each *d* value represents the magnitude of difference between the means for primary related occupations and non-primary occupations within a given target O*NET-SOC, and Table 10 and Figure 1 describe the distributions of these *d* values across all target O*NET-SOCs for which data were available to quantify such a difference. Each *d* value was computed as the difference between the independent means of primary related occupations and non-primary occupations within a given target O*NET-SOC (see Step 2 from our three-step analysis process). The average *d* value for relatedness scores was 2.92, while the absolute values of mean *d* values for the external criteria ranged from .66 (median salary absolute difference; negative) to 1.38 (interest profile correlation; positive). As an example of how to interpret these results, the mean *d* value for interest profile correlations indicates that, across the 874 target O*NET-SOCs with interest profile data, the average magnitude of difference between the profile correlations for primary related occupations and non-primary occupations was 1.38 standard deviations within target O*NET-SOCs.

Table 10. Summary of Cohen's d Values Comparing the Means of Criterion Variables Between Primary and Non-Primary Related Occupations

Criterion	*k*	*M*	*SD*	*Min*	5^{th} *%ile*	*Mdn*	95^{th} *%ile*	*Max*
Relatedness Score*	923	2.92	0.55	2.03	2.19	2.87	3.93	5.55
Skill Importance Profile Correlation	873	1.12	0.40	-0.83	0.54	1.04	1.79	2.82
Ability Importance Profile Correlation	873	1.10	0.41	-0.23	0.56	0.96	1.84	2.54
Interest Profile Correlation	874	1.38	0.52	-0.16	0.73	1.19	2.36	3.25
Work Style Profile Correlation	873	0.96	0.42	-0.88	0.18	1.02	1.54	2.20
Work Value Profile Correlation	874	0.90	0.45	-0.96	0.04	0.97	1.57	1.85
Median Salary Absolute Difference	923	-0.66	0.74	-4.45	-1.49	-0.63	0.11	3.91
Job Zone Absolute Difference	923	-0.95	0.39	-1.60	-1.59	-1.01	-0.26	0.58

Note. *k* = Number of target data-level O*NET-SOCs with data available for the criterion.

* Relatedness scores are not external criteria; they are included here as a point of comparison because they are the variable most directly related to the distinction between primary and non-primary related occupations.

All the mean *d* values were in directions that support the quality of the primary related occupations: the mean *d* values for profile correlations were positive, while the mean *d* values for median salary absolute differences and job zone absolute differences were negative, both of which indicate that the primary related occupations were more similar to their target O*NET-SOCs than were the rest of the occupations. Out of the all the criteria, only median salary absolute differences had a distribution of *d* values in which the 5th and 95th percentiles did not have the same sign, suggesting that similarity in salary between primary related occupations and target O*NET-SOCs was the least generalizable effect.

Figure 1. Distributions of Cohen's d Values Comparing the Means of Criterion Variables Between Primary and Non-Primary Related Occupations

Note. Solid vertical gray lines are reference lines for mean differences of zero. Solid vertical red lines indicate median effects and dashed vertical red lines indicate 5th and 95th percentiles.

Variance Reduction Analyses

Our analyses of the mean differences between primary related occupations and non-primary occupations showed clear distinctions between these sets of occupations on a variety of external criteria. Another way to evaluate the sets of primary related occupations is to consider how efficiently these sets of occupations reduce the variance of the criteria when compared to the complete distribution of potentially related occupations. In other words, it can be informative to evaluate how homogenous the primary related occupations are in comparison to the complete distribution of occupations. For example, if all the potentially related occupations for a target O*NET-SOC had a variance of .50 on a criterion while the primary related occupations had a variance of .10, we could say that the process by which the primary related occupations were identified reduced variance on the criterion by 80% (a variance reduction ratio of .80 = 1 - .10 / .50). This finding would indicate that, by demonstrating much less variance in scores than the complete distribution of occupations, the primary related occupations are quite homogenous compared to a random sampling of occupations. A large variance reduction ratio, combined with a large Cohen's *d* value, would indicate that primary related occupations are both homogenous and different from non-primary occupations.

Table 11 summarizes variance reduction ratios (VRRs) for our set of criteria, and Figure 2 depicts the distributions of these effects graphically. The mean VRRs ranged from .62 (median salary absolute difference) to .87 (ability importance profile correlation), indicating that, across target O*NET-SOCs, the primary related occupations tended to be much less variable than we would find in a random sampling of occupations. Although one might intuitively expect VRRs to range from 0 to 1, it is possible for these values to be negative, particularly if the primary related occupations are associated with outlier values that inflate their variance. Of the seven external criteria, only the median salary absolute difference criterion had a 5th percentile VRR value below zero; this means that, much like our *d* value analyses, only the VRRs for median salary absolute differences did not reliably generalize across target O*NET-SOCs. Other 5th percentile VRRs were close to zero (e.g., for work style profile correlations and work value profile correlations) but were still positive and supported widespread and reliably positive VRR effects.

Table 11. Summary of Criterion Variance-Reduction Ratios for Primary Related Occupations

Criterion	*k*	*M*	*SD*	*Min*	5th %ile	*Mdn*	95th %ile	*Max*
Relatedness Score*	923	0.86	0.19	-0.85	0.52	0.92	0.99	1.00
Skill Importance Profile Correlation	873	0.84	0.32	-2.03	0.13	0.96	1.00	1.00
Ability Importance Profile Correlation	873	0.87	0.27	-2.64	0.41	0.96	1.00	1.00
Interest Profile Correlation	874	0.83	0.27	-1.48	0.26	0.94	1.00	1.00
Work Style Profile Correlation	873	0.66	0.31	-1.21	0.01	0.74	0.96	0.99
Work Value Profile Correlation	874	0.68	0.31	-1.09	0.04	0.78	0.99	1.00
Median Salary Absolute Difference	923	0.62	0.75	-5.22	-0.92	0.86	0.99	1.00
Job Zone Absolute Difference	923	0.72	0.24	-0.99	0.34	0.80	1.00	1.00

Note. *k* = Number of target data-level O*NET-SOCs with data available for the criterion.

* Relatedness scores are not external criteria; they are included here as a point of comparison because they are the variable most directly related to the distinction between primary and non-primary related occupations.

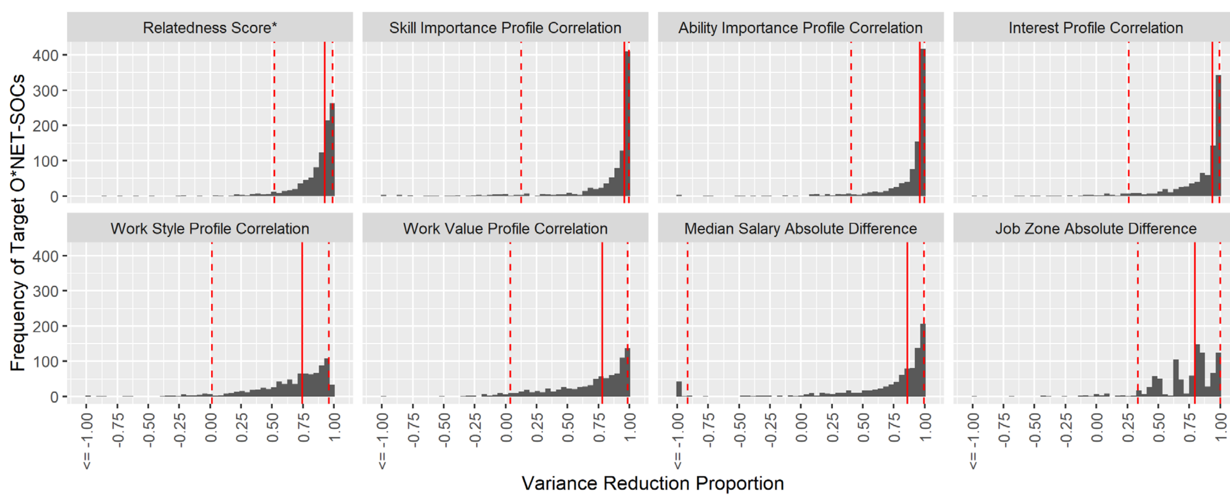

Figure 2. Distributions of Criterion Variance-Reduction Ratios for Primary Related Occupations

Note. Solid vertical red lines indicate median effects and dashed vertical red lines indicate 5th and 95th percentiles.

Collectively, our analyses of mean difference and variance-reduction ratios for external criteria provide strong evidence that the primary related occupations are distinct from non-primary occupations or random samples of occupations. These results support the usage of our relatedness composite (with expert review) to identify sets of primary related occupations that are highly similar to target O*NET-SOCs in terms of KSAO profiles, job zones, and, to a somewhat lesser extent, median salary.

Summary of Related Occupations Work Products and Future Updates

As noted earlier in the report, our efforts yielded two main work products for O*NET:

- **Operational Related Occupations Matrix**: A matrix showing 10 primary related occupations for each of the 923 data-level O*NET-SOCs (divided into two sets of five; the "Primary-Short" set contains the highest-priority relations to display, followed by the "Primary-Long" set) as well as 10 supplemental related occupations per data-level O*NET-SOC. Within each target O*NET-SOC's set of related occupations, the related O*NET-SOCs are assigned index values from 1–20 that reflects the order of O*NET-SOC mappings based on a combination of empirical similarity and expert review.
- **Related Occupations Research Dataset**: A complete record of quantitative occupational similarity information describing how each data-level O*NET-SOC relates to each of the other 922 data-level O*NET-SOCs. This dataset includes task and DWA WB-OSM composites, knowledge importance cosines, alternate title cosines, empirical relatedness scores, and the rank-order of empirical relatedness scores for each of the 923 data-level O*NET-SOCs. This dataset also includes all the information from the Operational Related Occupations Matrix.

Appendix C contains data dictionaries/codebooks for both work products.

Future Updates

One request the Center made of HumRRO at the outset of this work was to ensure the process for updating the related occupations data would be as streamlined as possible facilitate future updates. As such, as part of this effort, we developed *R* code that can be applied to updated versions of the O*NET database on a yearly basis to produced refreshed versions of both work products above. The code accomplishes the following:

- Ingests and processes O*NET task, DWA, knowledge, and alternate title data for use in calculating the final relatedness composite.
- Calculates the final relatedness composite scores for each pair of data-level occupations for which the aforementioned data are available.
- Compiles draft updated versions of the Operational Related Occupations Matrix and Related Occupations Dataset based on the final relatedness composite scores above for review by the Center review prior to release on O*NET sites.
- Compares the primary related occupations for each target occupations in the draft updated Operational Related Occupations Matrix to the current published Operational Related Occupations Matrix and flags target occupations where primary related occupations have changed. Unlike the initial creation of the Operational Related Occupations Matrix summarized in this report, the Center will only need to review results for the subset of target occupations where primary related occupations have changed.

References

Allen, M.T., Waugh, G.W., Shaw, M., Tsacoumis, S., et al. (2012). *The development and evaluation of a new O*NET Related Occupations Matrix (ROM)* (FR-12-29). Alexandria, VA: Human Resources Research Organization. https://www.onetcenter.org/reports/Related.html

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder. *ArXiv:1803.11175 [Cs]*. http://arxiv.org/abs/1803.11175

Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment*. Minneapolis, MN: University of Minnesota Press.

Drewes, D.W., Tarantino, J.W., Atkins, S.G., & Paige, B.A. (1999). *Development of the O*NET Related Occupations Matrix (ROM)*. Raleigh, NC: National Center for O*NET Development.

Equal Employment Opportunity Commission et al. (1978, August). Uniform Guidelines on Employee Selection Procedures. *Federal Register*, *43*(166) 38290-38315.

Gregory, C., Lewis, P., Frugoli, P., & Nallin, A. (2019). *Updating the O*NET®-SOC taxonomy: Incorporating the 2018 SOC structure*. Raleigh, NC: National Center for O*NET Development. https://www.onetcenter.org/reports/Taxonomy2019.html

Hansen, J. I., & Wiernik, B. M. (2018). Work preferences: Vocational interests and values. In D. S. Ones, N. Anderson, C. Viswesvaran, & H. K. Sinangil (Eds.), *The SAGE handbook of industrial, work, & organizational psychology: Personnel psychology and employee performance* (pp. 408-448). Los Angeles, CA: Sage Reference.

Harvey, R. J. (1991). Job analysis. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 71-163). Palo Alto, CA: Consulting Psychologists Press.

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

Lee, Y.-Y., Ke, H., Huang, H.-H., & Chen, H.-H. (2016). Less is more: Filtering abnormal dimensions in GloVe. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 71–72. https://doi.org/10.1145/2872518.2889381

Mototwidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*, 71-83.

National Center for O*NET Development (O*NET; 2015). *Ranking detailed work activities (DWAs) within O*NET occupational profiles*. Raleigh, NC: National Center for O*NET Development. https://www.onetcenter.org/reports/DWA_Ranking.html

National Center for O*NET Development. (2021a). O*NET 25.3 database. https://www.onetcenter.org/dictionary/25.3/excel/

National Center for O*NET Development. (2021b). O*NET 26.1 database. https://www.onetcenter.org/dictionary/26.1/excel/

National Center for O*NET Development. (2022). O*NET 26.2 database. https://www.onetcenter.org/dictionary/26.2/excel/

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Reimers, N., & Gurevych, I. (2019). Sentence-Bert: Sentence embeddings using Siamese BERT-networks. *ArXiv:1908.10084 [Cs]*. http://arxiv.org/abs/1908.10084

Rivkin, D., & Craven, D. E. (2021). *Procedures for O*NET job zone assignment: Updated to include procedures for developing preliminary job zones for new O*NET-SOC occupations*. Raleigh, NC: National Center for O*NET Development. https://www.onetcenter.org/reports/JobZoneProcedureUpdate.html

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Society for Industrial and Organizational Psychology (2018). *Principles for the validation and use of personnel selection procedures*, 5th edition. Bowling Green, OH: SIOP.

U.S. Bureau of Labor Statistics. (2020). *Overview of BLS wage data by area and occupation*. Washington, DC: U.S. Bureau of Labor Statistics. https://www.bls.gov/bls/blswage.htm

Appendix A: Mean Correlations Among Criteria and WB-OSM Variables Across the 923 Data-Level O*NET-SOCs

Variable Type	Index	Variable	Mean	SD	1	2	3	4	5	6	7	8	9	10	11
Criteria	1	KSA Distance Composite	0.00	0.75											
	2	Work Styles Shape	0.49	0.23	-.32										
	3	Interests Shape	0.17	0.47	-.55	.31									
	4	Work Values Shape	0.11	0.54	-.42	.28	.36								
	5	Abs. Med. Salary Diff.	35578.42	28129.35	.28	-.17	-.14	-.30							
Description Cosines	6	TF	0.15	0.07	-.13	.06	.10	.06	-.03						
	7	TF-IDF	0.01	0.02	-.26	.12	.20	.13	-.07	.47					
	8	GloVe (Unweighted)	0.31	0.11	-.51	.21	.44	.22	-.11	.37	.49				
	9	GloVe (TF-IDF Weighted)	0.18	0.12	-.52	.22	.47	.22	-.11	.26	.47	.95			
	10	USE-DAN	0.10	0.07	-.39	.20	.35	.21	-.12	.41	.49	.65	.64		
	11	SBERT	0.30	0.10	-.47	.23	.47	.23	-.14	.31	.45	.74	.76	.68	
Unweighted Task Cosines	12	TF	0.13	0.03	-.13	.08	.12	.09	-.07	.15	.12	.16	.16	.17	.20
	13	TF-IDF	0.01	0.01	-.49	.21	.35	.24	-.15	.22	.40	.51	.53	.44	.54
	14	GloVe (Unweighted)	0.31	0.07	-.64	.25	.50	.28	-.18	.18	.31	.66	.68	.44	.61
	15	GloVe (TF-IDF Weighted)	0.19	0.08	-.65	.26	.52	.28	-.16	.18	.31	.68	.71	.45	.63
	16	USE-DAN	0.11	0.04	-.57	.27	.48	.29	-.17	.18	.33	.56	.59	.53	.63
	17	SBERT	0.29	0.06	-.59	.26	.52	.27	-.16	.18	.32	.62	.66	.49	.72
Weighted Task Cosines	18	TF	0.13	0.03	-.13	.08	.12	.09	-.07	.15	.12	.16	.16	.17	.21
	19	TF-IDF	0.01	0.01	-.49	.21	.36	.24	-.16	.22	.40	.51	.53	.44	.54
	20	GloVe (Unweighted)	0.31	0.07	-.64	.25	.50	.28	-.18	.18	.31	.66	.69	.45	.61
	21	GloVe (TF-IDF Weighted)	0.19	0.08	-.65	.26	.52	.28	-.16	.18	.32	.68	.72	.46	.63
	22	USE-DAN	0.11	0.04	-.57	.27	.48	.29	-.17	.18	.33	.56	.59	.53	.63
	23	SBERT	0.29	0.06	-.59	.26	.52	.27	-.16	.18	.32	.62	.66	.49	.72
Unweighted DWA Cosines	24	TF	0.09	0.02	-.31	.16	.34	.20	-.16	.13	.20	.30	.30	.32	.37
	25	TF-IDF	0.02	0.01	-.52	.25	.45	.26	-.18	.17	.35	.50	.52	.44	.54
	26	GloVe (Unweighted)	0.28	0.06	-.61	.26	.52	.30	-.17	.17	.29	.61	.63	.43	.59
	27	GloVe (TF-IDF Weighted)	0.20	0.06	-.62	.25	.51	.28	-.14	.17	.30	.64	.66	.42	.59
	28	USE-DAN	0.15	0.03	-.58	.27	.51	.30	-.19	.17	.31	.54	.56	.47	.57
	29	SBERT	0.31	0.05	-.58	.27	.54	.29	-.18	.17	.29	.56	.59	.45	.64
Weighted DWA Cosines	30	TF	0.09	0.03	-.31	.17	.35	.20	-.14	.15	.21	.30	.31	.32	.37
	31	TF-IDF	0.02	0.01	-.51	.25	.44	.25	-.17	.17	.35	.50	.52	.44	.53
	32	GloVe (Unweighted)	0.28	0.06	-.60	.28	.53	.29	-.17	.18	.29	.61	.63	.44	.60
	33	GloVe (TF-IDF Weighted)	0.20	0.07	-.61	.26	.52	.27	-.14	.18	.31	.64	.66	.43	.60
	34	USE-DAN	0.15	0.04	-.57	.27	.50	.29	-.18	.18	.31	.54	.55	.47	.57
	35	SBERT	0.31	0.06	-.57	.28	.54	.27	-.18	.18	.30	.56	.59	.46	.65

Notes. Abs. Med. Salary Diff. = Absolute Value of Median Salary Difference. TF = term frequency. TF-IDF = term frequency multiplied by inverse document frequency. GloVe (Unweighted) = GloVe embeddings computed by taking the simple, unweighted average of word-level embeddings to generate aggregated sentence-level embeddings. GloVe (TF-IDF Weighted) = GloVe embeddings computed by taking the TF-IDF weighted average of word-level embeddings to generate aggregated sentence-level embeddings. USE-DAN = sentence-level embeddings generated using Google's Universal Sentence Encoder (USE) with Deep Averaging Network (DAN). SBERT = sentence-level embeddings generated using the Sentence BERT model in Python. Values were calculated within each O*NET-SOC and then averaged across O*NET-SOCs. For example, intercorrelations were calculated for O*NET-SOC 1 vs. all other O*NET-SOCs, O*NET-SOC 2 vs. all other O*NET-SOCs, etc., and then averaged. Correlations are color-coded by direction and magnitude.

Variable Type	Index	Variable	12	13	14	15	16	17	18	19	20	21	22	23
Unweighted Task Cosines	12	TF												
	13	TF-IDF	.44											
	14	Glove (Unweighted)	.31	.72										
	15	Glove (TF-IDF Weighted)	.23	.70	.98									
	16	USE-DAN	.42	.73	.78	.78								
	17	SBERT	.29	.71	.86	.88	.84							
Weighted Task Cosines	18	TF	1.00	.44	.31	.23	.42	.29						
	19	TF-IDF	.44	1.00	.72	.70	.73	.71	.44					
	20	Glove (Unweighted)	.31	.72	1.00	.98	.79	.86	.31	.72				
	21	Glove (TF-IDF Weighted)	.23	.70	.98	1.00	.78	.88	.23	.70	.98			
	22	USE-DAN	.42	.73	.78	.78	1.00	.84	.42	.73	.78	.78		
	23	SBERT	.29	.71	.86	.88	.84	1.00	.29	.71	.86	.88	.84	
Unweighted DWA Cosines	24	TF	.18	.36	.33	.34	.42	.40	.18	.36	.34	.34	.42	.41
	25	TF-IDF	.18	.59	.59	.61	.60	.62	.18	.59	.60	.61	.60	.62
	26	Glove (Unweighted)	.16	.57	.78	.79	.65	.74	.16	.57	.78	.79	.65	.74
	27	Glove (TF-IDF Weighted)	.14	.58	.80	.82	.64	.75	.14	.58	.80	.82	.64	.75
	28	USE-DAN	.18	.56	.65	.66	.66	.69	.18	.56	.65	.67	.66	.69
	29	SBERT	.20	.57	.72	.74	.68	.80	.20	.58	.72	.74	.68	.80
Weighted DWA Cosines	30	TF	.18	.36	.34	.34	.41	.40	.18	.37	.34	.34	.42	.40
	31	TF-IDF	.18	.60	.59	.60	.60	.62	.18	.60	.59	.60	.60	.62
	32	Glove (Unweighted)	.17	.57	.77	.78	.65	.74	.17	.57	.78	.79	.65	.74
	33	Glove (TF-IDF Weighted)	.15	.59	.80	.82	.65	.76	.15	.59	.80	.82	.65	.76
	34	USE-DAN	.18	.56	.65	.66	.66	.68	.18	.57	.65	.66	.66	.69
	35	SBERT	.20	.58	.72	.73	.69	.80	.20	.58	.72	.73	.69	.81

Variable Type	Index	Variable	24	25	26	27	28	29	30	31	32	33	34
Unweighted DWA Cosines	24	TF											
	25	TF-IDF	.64										
	26	Glove (Unweighted)	.58	.77									
	27	Glove (TF-IDF Weighted)	.42	.75	.96								
	28	USE-DAN	.73	.82	.86	.81							
	29	SBERT	.58	.78	.89	.86	.87						
Weighted DWA Cosines	30	TF	.93	.61	.56	.41	.69	.56					
	31	TF-IDF	.61	.95	.73	.72	.79	.75	.64				
	32	Glove (Unweighted)	.57	.75	.97	.93	.84	.87	.60	.75			
	33	Glove (TF-IDF Weighted)	.42	.74	.94	.97	.79	.85	.44	.74	.96		
	34	USE-DAN	.69	.79	.83	.78	.96	.84	.72	.80	.86	.80	
	35	SBERT	.57	.76	.86	.83	.85	.97	.58	.77	.88	.86	.86

Appendix B: Cumulative Frequency Tables Summarizing Rates at Which Related Occupations Exceed the Bookmarking Threshold Across Old and New Related Occupations Matrices

*Table B.1. Cumulative Frequency Distribution for Proportions of Related O*NET-SOCs that Exceeded the Bookmarking Threshold in the Career Starters Matrix*

Proportion of Related Occupations Above Threshold	Non-Cumulative Summary of Target O*NET-SOCs		Cumulative Summary of Target O*NET-SOCs With Proportions Less Than or Equal to Row Proportion		Cumulative Summary of Target O*NET-SOCs With Proportions Greater Than or Equal to Row Proportion	
	k	%	*k*	%	*k*	%
0.0	70	9.01	70	9.01	777	100.00
0.1	70	9.01	140	18.02	707	90.99
0.2	75	9.65	215	27.67	637	81.98
0.3	75	9.65	290	37.32	562	72.33
0.4	69	8.88	359	46.20	487	62.68
0.5	76	9.78	435	55.98	418	53.80
0.6	89	11.45	524	67.44	342	44.02
0.7	66	8.49	590	75.93	253	32.56
0.8	88	11.33	678	87.26	187	24.07
0.9	61	7.85	739	95.11	99	12.74
1.0	38	4.89	777	100.00	38	4.89

*Table B.2. Cumulative Frequency Distribution for Proportions of Related O*NET-SOCs that Exceeded the Bookmarking Threshold in the Career Changers Matrix*

Proportion of Related Occupations Above Threshold	Non-Cumulative Summary of Target O*NET-SOCs		Cumulative Summary of Target O*NET-SOCs With Proportions Less Than or Equal to Row Proportion		Cumulative Summary of Target O*NET-SOCs With Proportions Greater Than or Equal to Row Proportion	
	k	%	*k*	%	*k*	%
0.0	18	2.32	18	2.32	777	100.00
0.1	48	6.18	66	8.49	759	97.68
0.2	34	4.38	100	12.87	711	91.51
0.3	59	7.59	159	20.46	677	87.13
0.4	44	5.66	203	26.13	618	79.54
0.5	73	9.40	276	35.52	574	73.87
0.6	70	9.01	346	44.53	501	64.48
0.7	83	10.68	429	55.21	431	55.47
0.8	82	10.55	511	65.77	348	44.79
0.9	59	7.59	570	73.36	266	34.23
1.0	207	26.64	777	100.00	207	26.64

*Table B.3. Cumulative Frequency Distribution for Proportions of Related O*NET-SOCs that Exceeded the Bookmarking Threshold in the Empirical Top-10 Matrix*

Proportion of Related Occupations Above Threshold	Non-Cumulative Summary of Target O*NET-SOCs		Cumulative Summary of Target O*NET-SOCs With Proportions Less Than or Equal to Row Proportion		Cumulative Summary of Target O*NET-SOCs With Proportions Greater Than or Equal to Row Proportion	
	k	%	*k*	%	*k*	%
0.0	4	0.43	4	0.43	923	100.00
0.1	2	0.22	6	0.65	919	99.57
0.2	16	1.73	22	2.38	917	99.35
0.3	14	1.52	36	3.90	901	97.62
0.4	9	0.98	45	4.88	887	96.10
0.5	14	1.52	59	6.39	878	95.12
0.6	15	1.63	74	8.02	864	93.61
0.7	8	0.87	82	8.88	849	91.98
0.8	12	1.30	94	10.18	841	91.12
0.9	11	1.19	105	11.38	829	89.82
1.0	818	88.62	923	100.00	818	88.62

*Table B.4. Cumulative Frequency Distribution for Proportions of Related O*NET-SOCs that Exceeded the Bookmarking Threshold in the Empirical Top-10 Matrix (Limited to the O*NET-SOCs included in the Career Starters and Career Changers Matrices)*

Proportion of Related Occupations Above Threshold	Non-Cumulative Summary of Target O*NET-SOCs		Cumulative Summary of Target O*NET-SOCs With Proportions Less Than or Equal to Row Proportion		Cumulative Summary of Target O*NET-SOCs With Proportions Greater Than or Equal to Row Proportion	
	k	%	*k*	%	*k*	%
0.0	3	0.39	3	0.39	777	100.00
0.1	1	0.13	4	0.51	774	99.61
0.2	11	1.42	15	1.93	773	99.49
0.3	13	1.67	28	3.60	762	98.07
0.4	9	1.16	37	4.76	749	96.40
0.5	14	1.80	51	6.56	740	95.24
0.6	13	1.67	64	8.24	726	93.44
0.7	6	0.77	70	9.01	713	91.76
0.8	12	1.54	82	10.55	707	90.99
0.9	11	1.42	93	11.97	695	89.45
1.0	684	88.03	777	100.00	684	88.03

*Table B.5. Cumulative Frequency Distribution for Proportions of Related O*NET-SOCs that Exceeded the Bookmarking Threshold in the Final Operational Related Occupations Matrix (Limited to the O*NET-SOCs included in the Career Starters and Career Changers Matrices)*

Proportion of Related Occupations Above Threshold	Non-Cumulative Summary of Target O*NET-SOCs		Cumulative Summary of Target O*NET-SOCs With Proportions Less Than or Equal to Row Proportion		Cumulative Summary of Target O*NET-SOCs With Proportions Greater Than or Equal to Row Proportion	
	k	%	*k*	%	*k*	%
0.0	3	0.39	3	0.39	777	100.00
0.1	1	0.13	4	0.51	774	99.61
0.2	12	1.54	16	2.06	773	99.49
0.3	16	2.06	32	4.12	761	97.94
0.4	14	1.80	46	5.92	745	95.88
0.5	15	1.93	61	7.85	731	94.08
0.6	14	1.80	75	9.65	716	92.15
0.7	14	1.80	89	11.45	702	90.35
0.8	21	2.70	110	14.16	688	88.55
0.9	39	5.02	149	19.18	667	85.84
1.0	628	80.82	777	100.00	628	80.82

Appendix C: Data Dictionaries/Codebooks for the Operational Related Occupations Matrix and Related Occupations Research Dataset

Table C.1. Data Dictionary/Codebook for the Operational Related Occupations Matrix

Variable Name	Variable Type	Variable Description
O*NET-SOC Code	Nominal String	O*NET-SOC codes for target occupations.
Title	Nominal String	O*NET titles for target occupations.
Related O*NET-SOC Code	Nominal String	O*NET-SOC codes for related occupations.
Related Title	Nominal String	O*NET titles for related occupations.
Relatedness Tier	Ordinal String	Ordinal categories indicating level of relatedness after expert review. Determined separately within each target O*NET-SOC. Used to determine which related occupations to display on web resources. - Primary-Short = Five most strongly related occupations after expert review. - Primary-Long = 6th to 10th most strongly related occupations after expert review. - Supplemental = 11th to 20th most strongly related occupations after expert review.
Index	Integer	Ordering of related occupations after expert review. Determined separately within each target O*NET-SOC.

Table C.2. Data Dictionary/Codebook for the Related Occupations Research Dataset

Variable Name	Variable Type	Variable Description
O*NET-SOC Code	Nominal String	O*NET-SOC codes for target occupations.
Title	Nominal String	O*NET titles for target occupations.
Related O*NET-SOC Code	Nominal String	O*NET-SOC codes for related occupations.
Related Title	Nominal String	O*NET titles for related occupations.
Relatedness Tier	Ordinal String	Ordinal categories indicating level of relatedness after expert review. Determined separately within each target O*NET-SOC. Used to determine which related occupations to display on web resources. - Primary-Short = Five most strongly related occupations after expert review. - Primary-Long = 6th to 10th most strongly related occupations after expert review. - Supplemental = 11th to 20th most strongly related occupations after expert review. - N/A = Not categorized.
Index	Integer	Ordering of related occupations after expert review. Determined separately within each target O*NET-SOC. Values are only displayed for related occupations that were included in the Operational Related Occupations Matrix.
Work-Based Occupational Similarity	Numeric	Simple average of SBERT-based WB-OSMs computed using (a) task statements and (b) detailed work activities (DWAs).
Knowledge Cosine	Numeric	Cosine similarity between occupation's profiles of importance ratings for O*NET's 33 knowledge domains.
Alternate Titles Cosine	Numeric	Cosine similarity between the TF-IDF-weighted GloVe embeddings for occupations' alternate titles.
Relatedness Score	Numeric	Final quantitative metric for determining occupational similarity/relatedness.
Empirical Relatedness Rank	Integer	Rank ordering of related occupations based on descending values of Relatedness Scores within each target O*NET-SOC.