HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2023 No. 147

Using Machine Learning to Develop Occupational Interest Profiles and High-Point Codes for the O*NET System

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:	Dan J. Putka, HumRRO Jeffrey A. Dahlke, HumRRO Maura I. Burke, HumRRO James Rounds, University of Illinois Phil Lewis, National Center for O*NET Development	**Date:**	November 7, 2023

Using Machine Learning to Develop Occupational Interest Profiles and High-Point Codes for the O*NET System

Table of Contents

Table of Contents (Continued)

List of Tables

Table of Contents (Continued)

List of Figures

Using Machine Learning to Develop Occupational Interest Profiles and High-Point Codes for the O*NET System

Introduction

The Occupational Information Network (O*NET) is a comprehensive system developed by the U.S. Department of Labor that provides information for over 900 occupations within the U.S. economy. This information is maintained in a database (National Center for O*NET Development, 2023). 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. The purpose of this project was to develop updated Occupational Interest Profiles (OIPs) and high-point codes for the 923 data-level occupations included within the O*NET-SOC 2019 taxonomy (Gregory et al., 2019) and to offer a streamlined process for future updating of interest data in the O*NET System.

Vocational interest information is an important part of the O*NET Program's support of educational planning, career exploration, career guidance, job search, and organizational placement (Rounds et al., 2021). The O*NET Content Model defines interest information compatible with Holland's RIASEC model of personality types and work environments (Holland, 1997) and was recently expanded to include basic interests related to each RIASEC dimension (Rounds et al., 2023). The RIASEC model serves as a foundation for interest information in the O*NET System due to its extensive use in applied settings and research. The model is familiar to and preferred by vocational counselors and employment program professionals/directors (e.g., One Stop Centers, Employment Security Offices, Workforce Development Centers). The move towards and emphasis on self-assessment in career exploration also necessitated the selection of a model that was easy for end users to understand and use, and the RIASEC model fulfills that need.

Within the RIASEC model, six interest categories are used to describe the work environment of occupations: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional (i.e., RIASEC). Table A.1 in Appendix A provides descriptions of each RIASEC interest category as described in the O*NET Content Model. Corresponding OIPs and high-point codes are published within the O*NET database and O*NET web services. An OIP consists of six numerical scores indicating how descriptive each RIASEC dimension is of an O*NET-SOC occupation. In addition, a profile of one to three high-point codes indicates which RIASEC dimensions are most descriptive of an O*NET-SOC occupation. The number of high-point codes offered for an O*NET occupation depends on the number that meets a minimum degree of descriptiveness for the O*NET-SOC occupation. Details on how OIPs and high-point codes have historically been developed for occupations in O*NET are detailed in Rounds et al. (1999), Rounds et al. (2008), and Rounds et al. (2013). Interest data for O*NET-SOC occupations is disseminated through various O*NET websites (O*NET OnLine; My Next Move; My Next Move for Veterans; Mi Proximo Paso), available through the O*NET database and web services (National Center for O*NET Development, 2023), and play a critical role in facilitating career exploration processes in conjunction with O*NET's Interest Profiler assessment to identify O*NET-SOC occupations that may align with an individual's vocational interests (Gregory & Lewis, 2016; Rounds et al., 2021).

The O*NET Program's recent transition to the O*NET-SOC 2019 taxonomy structure necessitates the development and dissemination of Interest Profiles (OIPs) and high-point codes for the 923 data-level occupations included in that taxonomy. Additionally, the last major updates to interest data in the O*NET database occurred in 2008 and 2013 (Rounds et al., 2008; Rounds et al., 2013). Thus, there is a need to update the interest data in the O*NET database to fully cover data-level occupations in the O*NET-SOC 2019 taxonomy and account for changes that may have occurred in occupations over the past decade.

Updating Occupational Interest Profiles and High-Point Codes

The current effort aimed to develop updated OIPs and high-point codes for the 923 data-level occupations included within the O*NET-SOC 2019 taxonomy and offer a streamlined process for future updating of interest data in O*NET. There are at least three approaches to developing OIPs for occupations, including (a) incumbent, (b) empirical, and (c) judgment-based approaches (Rounds et al., 1999). The incumbent approach involves assessing the interests of occupational incumbents and aggregating them to the occupation level.[1] The empirical approach involves using occupation-level data to make predictions about how descriptive an interest is of an occupation's work environment. Lastly, the judgment approach involves asking subject matter experts to provide judgments of how descriptive an interest is of an occupation's work environment.

There have been three major updates to the interest data in O*NET over time. The original OIP development work in 1999 explored and compared the use of one type of empirical approach versus a judgment approach to populating OIPs and high-point code data (Rounds et al., 1999). The empirical approach examined in the 1999 effort was found to have limited utility beyond identifying the most relevant interest for an occupation (i.e., the first high-point code). This led to the development of an approach using direct ratings of occupations by trained raters (i.e., a judgment approach). The raters made judgments based on standardized stimulus materials available from O*NET data at the time, which included occupation titles, descriptions, and core tasks. A subject matter expert (SME) then refined the high-point codes as needed following the rating exercise. The quality of the resulting data was supported with both structural and external validity evidence (see Rounds et al., 1999 for a summary). The judgment-based approach was also used in the 2008 and 2013 updates to OIPs and associated high-point codes (Rounds et al., 2008; Rounds et al., 2013).

While the previous judgment-based approaches produced high-quality OIPs and high-point codes, the resource-intensive nature of relying on trained raters leads to challenges in maintaining the currency of the database over time. Additionally, since the 1999 evaluation of the empirical approach to generating OIPs, there have been several developments that make empirical approaches far more viable options for O*NET to consider, most notably:

- The emergence of high-quality criterion data on which to train RIASEC prediction models (i.e., numeric RIASEC ratings for O*NET-SOCs provided by trained human raters; e.g., Rounds et al., 2008; Rounds et al., 2013).

[1] Though earlier championed by Holland (1997), the incumbent approach is problematic in that it does not provide a direct measure of how characteristic or descriptive each interest is of an occupation's work environment, but rather the interests of the individuals who work in an occupation. These represent two different targets of measurement (i.e., characteristics of individuals vs. characteristics of work in occupations). The purpose of the OIPs in O*NET is to provide interest profiles for occupations based on the work performed within those occupations, rather than the people who work in them.

- The emergence of methods for quantifying occupation text data (e.g., descriptions, task statements) that bear conceptual relevance to RIASEC dimensions and that could be used as inputs to predictive models (Dahlke & Putka, 2022; Dahlke et al., 2022; Putka et al., 2023).
- The emergence of supervised learning methods that can generate more generalizable predictions in the face of limited data to train and cross-validate models (e.g., regularized regression models that help avoid capitalization on chance and yield higher levels of cross-validity than traditional modeling methods; James et al. 2021).[2]

Overview of Development and Evaluation Approach

Given the objectives of this effort and lessons learned from past research, our technical approach to updating OIPs and high-point codes reflects a combination of empirical and judgment-based approaches and is based on the following assumptions:

- Interest ratings for occupations made by trained human raters are of high quality. This is evidenced by the analysis presented in past O*NET reports (Rounds et al., 1999; Rounds et al., 2008; Rounds et al., 2013).
- There are strong conceptual relations between what people do on a job (as captured in occupation descriptions and task statement lists) and the RIASEC interests that best describe work in that job. These conceptual relations manifest in empirical relations between (a) O*NET-SOC descriptions and tasks and (b) human ratings of RIASEC dimensions that have magnitudes rivaling the level of interrater reliability observed among trained human raters (Dahlke & Putka, 2022; Putka et al., 2023).
- It is possible to generate accurate RIASEC interest ratings (and associated high-point codes) based on predictive models that use inputs from other parts of the O*NET Content Model without engaging in an extensive data collection with trained human raters (Dahlke & Putka, 2022; Putka et al., 2023).
- The aforementioned predictive models can afford the Center with a semi-automated means for updating OIPs and associated high-point codes over time as O*NET-SOCs change and new O*NET-SOCs emerge.

With these assumptions as a foundation, we engaged in a multi-step process to develop and evaluate updated OIPs and high-point codes for the 923 data-level occupations included within the O*NET-SOC 2019 taxonomy. These steps included:

- Step 1: Acquiring and Preparing Data for Initial Modeling Work
- Step 2: Developing and Evaluating Initial RIASEC Prediction Models
- Step 3: Generating Preliminary OIPs and High-Point Codes
- Step 4: Identifying Occupations for Inclusion in Analyst-Expert Rating Data Collections
- Step 5: Collecting and Evaluating Analyst and Expert RIASEC Ratings

[2] Supervised learning can be viewed as a falling within the broader domain of statistical or machine learning, with a focus on developing and validating models for predicting an outcome or criterion of interest based on a set of inputs.

- Step 6: Refining and Evaluating Final RIASEC Prediction Models for Future Use
- Step 7: Finalizing OIPs and High-Point Codes for publication in the O*NET 28.1 database

We provide details on each step in the sections that follow. Lastly, we conclude this report with guidance for updating OIPs and high-point codes in future versions of the O*NET database based on the prediction models developed herein.

Step 1: Acquiring and Preparing Data for Initial Modeling Work

To develop and evaluate models of how descriptive each RIASEC dimension is of an occupation, it is ideal to have samples of occupations for which high-quality RIASEC ratings already exist. Fortunately, in the case of O*NET, such samples are available from the 2008 and 2013 OIP development efforts (Rounds et al., 2008; Rounds et al., 2013).[3] The 2008 and 2013 development efforts provided a critical foundation for developing and evaluating RIASEC prediction models for occupations in the O*NET-SOC 2019 taxonomy. Specifically, we used these datasets to better understand which types of prediction models tend to best align with RIASEC ratings from trained human raters.

As a first step, we acquired and prepared the following data from various O*NET databases to support the modeling work subsequently described under Step 2[4]:

- Occupation descriptions, occupation titles, task statements,[5] knowledge, skill, and ability (KSA) importance ratings, generalized work activity importance ratings, RIASEC interest ratings, and RIASEC high-point codes for:
  - The 812 occupations in the O*NET 13.0 Database (June 2008) with interest data (i.e., the database underlying the 2008 version of the O*NET RIASEC data).
  - The 908 occupations in the O*NET 14.0 Database (June 2009) with interest data (i.e., the database underlying the 2009 version of the O*NET RIASEC data).
  - The 974 occupations in the O*NET 18.0 Database (July 2013) with interest data (i.e., the database underlying the 2013 version of the O*NET RIASEC data).

We also obtained the following data from Dr. James Rounds, who led the 2008 and 2013 OIP development efforts:

- Disaggregated RIASEC interest ratings from individual raters underlying the 2008, 2009, and 2013 versions of the O*NET RIASEC data.

Lastly, we obtained text for RIASEC and basic interest items that we subsequently used in our modeling process (detailed in Step 2).

- RIASEC items from O*NET's Interest Profiler (IP) Short Form measure (Rounds et al., 2010, 2021).
- Basic interest items from the CABIN measure (Su et al., 2019) and "Illustrative Activities" for basic interests developed for the 2023 expansion of the interest domain of the O*NET Content Model (Rounds et al., 2023).

---

[3] Note, though human-based ratings from the 1999 OIP and high-point code development work described by Rounds et al. (1999) are available through ONET 3.0 Transitional Database (published in August 2000), those ratings were made on occupational units that pre-date the O*NET-SOC taxonomic structure and as such were not considered for this effort.

[4] Later in our research effort, we also examined other text data that became available in later versions of the O*NET database after O*NET 18.0 (e.g., alternate titles, detailed work activities, intermediate work activities). Under Step 6, we explain how we used these additional types of text to further refine our initial models.

[5] We used core task statements for occupations that differentiated between core and supplemental tasks, and all tasks for occupations that did not make that differentiation.

Constructing a Combined 2008-2013 Dataset for Initial Modeling

With the data above, we first constructed a combined dataset with information from the three O*NET databases noted (13.0, 14.0, and 18.0). We subsequently refer to this as the "2008-2013 Dataset."

The 2008-2013 Dataset consists of the 974 O*NET-SOCs with interest data from O*NET 18.0, but the inputs tied to those occupations (i.e., occupation descriptions, occupation titles, task statements, KSA importance ratings, generalized work activity importance ratings) were drawn from either the O*NET 13.0, 14,0, or 18.0 database depending on when in time the interest data for that occupation was obtained from trained raters as part of previous OIP development efforts. For example, if the interest ratings for an occupation reflected ratings gathered for O*NET 13.0, then we use input data from O*NET 13.0 for that occupation, as opposed to what inputs were provided for that occupation in O*NET 18.0. This ensured the inputs and interest ratings were in close temporal alignment; that is, the inputs largely reflected the state of the occupation as described by O*NET at the time the interest ratings were first published.

A challenge in aligning the data across different versions of the O*NET database noted above is that they were not all based on the same O*NET-SOC taxonomy. Specifically, O*NET 18.0 was based on O*NET-SOC 2010, O*NET 14.0 was based on O*NET-SOC 2009, and O*NET 13.0 was based on O*NET-SOC 2006.

To address this matter, we used the O*NET-SOC 2010 as the foundation for the 2008-2013 Dataset. The rationale was that the O*NET-SOC 2010 taxonomy provided the basis for O*NET 18.0, which afforded the latest and most complete set of interest data among the databases examined. Next, using published O*NET crosswalks available through the O*NET Resource Center (O*NET-SOC Taxonomy), the O*NET-SOC 2010 codes in O*NET 18.0 were linked to O*NET-SOCs in the earlier database versions examined (i.e., O*NET 14.0 and 13.0). We then used those linkages to pull in occupational information (e.g., occupational descriptions, tasks) from the database that was most proximal to the date that interest occupation was updated for that occupation as documented in O*NET 18.0.

For example, if an occupation in O*NET 18.0 had an interest publication date of 2013, then we simply used occupation information that was present in O*NET 18.0 for the 2008-2013 Dataset (n = 83). If an occupation in O*NET 18.0 had an interest publication date of 2009, then we would crosswalk the O*NET-SOC 2010 with O*NET-SOC 2009 and merge in occupational information for that occupation from O*NET 14.0 for the 2008-2013 Dataset (n = 95). Lastly, if an occupation in O*NET 18.0 had an interest publication date of 2008, then we would crosswalk the O*NET-SOC 2010 with O*NET-SOC 2006 and merge in occupational information for that occupation from O*NET 13.0 for the 2008-2013 Dataset (n = 796).

Given the focus of the current effort on using RIASEC ratings from the 2008-2013 Dataset as our primary criteria for building predictive models, we took the added step of attempting to ensure the quality of those criteria (i.e., the judgment-based RIASEC ratings). As such, we revisited the original, disaggregated RIASEC ratings (i.e., RIASEC ratings made by individual trained raters) and examined them for disagreement for each occupation. The original judgment-based ratings for each RIASEC dimension simply reflected averages of ratings from trained individual raters, with no consideration of the level of agreement among those raters *for a given occupation*. This could result in a source of inaccuracy in the ratings as an outlying rater could readily skew the average rating for an occupation. HumRRO obtained disaggregated interest rating information from Dr. Rounds, which informed the original development of OIPs for

each of the O*NET 18.0, 14.0, and 13.0 releases. Specifically, separate datasets containing individual-level interest ratings for a given release were shared, allowing us to merge that information with the 2008-2013 Dataset based on the crosswalks described above.

Using the data above, we developed criteria for flagging the 5,844 occupation-RIASEC combinations (974 occupations x 6 RIASEC dimensions) where disagreement among the three trained raters who provided the original ratings was considered meaningful. In consultation with Dr. Round's and the Center, it was agreed that occupations would be flagged based on the following rules:

- Rule 1: Range across raters was greater than or equal to four (on a one-to-seven rating scale) AND if two of the raters were less than or equal to one rating point away from each other. 143 out of 5,844 (2.4.%) of occupation-RIASEC combinations met this rule. The occupation-RIASEC combinations that met this rule spanned 125 of the 974 (12.8%) occupations in the 2008-2013 Dataset.
- Rule 2: Range across raters was greater than or equal to four AND if two of the raters were greater than one rating point away from each other. 65 out of 5,844 (1.1%) of occupation-RIASEC combinations met this rule. The occupation-RIASEC combinations that met this rule spanned 61 of the 974 (6.3%) occupations in the 2008-2013 Dataset.
- Rule 3: Range across raters was equal to three AND if two of the raters gave the same rating. 146 out of 5,844 (2.5%) of occupation-RIASEC combinations met this rule. The occupation-RIASEC combinations that met this rule spanned 134 of the 974 (13.7%) occupations in the 2008-2013 Dataset.

Based on the rules above, we made the following adjustments to the original RIASEC ratings before building predictive models in Step 2:

- Trimmed means (or exact value, if same) were used in instances where an occupation-RIASEC combination was flagged for Rule 1, where the outlying rater's rating was removed prior to calculating the trimmed mean.
- Expert ratings from Dr. Rounds were used in instances where an occupation-RIASEC was flagged for Rule 2.[6]
- The modal (i.e., agree upon) rating was used in instances where an occupation-RIASEC combination was flagged for Rule 3.

Note that if an occupation-RIASEC combination was not flagged based on one of the rules above, the mean rating across raters was used as the final rating for a given RIASEC dimension (i.e., the ratings published in the O*NET database from which the RIASEC data were drawn).

Given the high-point codes in the O*NET database are a function of the RIASEC ratings, we revisited whether high-point codes for the occupations impacted by the rules above necessitated changes to their high-point codes. This evaluation revealed the need for us to

[6] HumRRO enlisted Dr. Rounds to provide expert judgment on occupations where past, individual rater information existed and flagged under Rule 2. Dr. Rounds was provided an Excel based rating booklet consisting of: (a) ratings instructions (see Appendix B), (b) a ratings sheet, (c) O*NET occupational descriptions, and (d) associated task statements. After completing the re-rating exercise, we considered Dr. Rounds' ratings as the new final rating associated with a given occupation-RIASEC combination for purposes of building prediction models.

update high-point codes for 27 of 974 (2.7%) occupations and 214 of 2,922 (7.3%) occupation-RIASEC high-point code combinations (974 occupations x 3 high-point codes) given changes to ratings based on the rules above.[7]

In sum, the final 2008-2013 Dataset was constructed such that the original RIASEC ratings were retained for an occupation if that occupation was not flagged for Rules 1, 2, or 3. Trimmed means or modal ratings were used for occupation-RIASEC combinations flagged based on Rules 1 and 3, respectively. Dr. Round's new ratings were for occupation-RIASEC combinations flagged based on Rule 2. This process resulted in a complete and final dataset with interest information for 974 occupations, which was subsequently used for the development and evaluation of initial RIASEC prediction models.

[7] 214 is the total of new RIASEC high-point codes when no tied situations occurred for the first position, first and second position, and finally, the first, second and third positions.

Step 2: Developing and Evaluating Initial RIASEC Prediction Models

As alluded to above, an important element of the current effort was conducting analyses to understand which types of prediction models are best for predicting trained human raters' RIASEC ratings. As such, our focus in this initial modeling step was on developing and evaluating a large set of potential models with the aim of what works best for maximizing prediction. During Step 2, we were not focused on models that would be sensitive to practical considerations (e.g., how soon certain model inputs would become available for use after a new O*NET-SOC is introduced into the O*NET database). We factored in practical considerations during later steps in our development effort (see Step 6 in this report). Additionally, focusing on identifying models that offered the best levels of prediction during Step 2—regardless of practicality—afforded us important benchmarks for later steps. Specifically, we used the best-performing models from Step 2 as benchmarks for determining how much predictive value would be lost (if any) were we to adopt a model that balanced both prediction and practical considerations for implementation in future versions of O*NET.

In light of our initial focus on prediction, our activities in this step included the following:

1. Preparing text for use in prediction models. As alluded to above, text can be quantified in different ways for use in prediction models. We quantified text from occupation descriptions, occupation titles, task statements, and interest items (IP and CABIN) using various methods, as needed, to support the models.

2. Developing specifications for 17 different models for predicting numeric RIASEC descriptiveness ratings for O*NET occupations. Models were differentiated by the types of O*NET data used as the basis for generating predictions (e.g., occupation descriptions, titles, task statements, importance ratings for KSAs) and how text-based inputs were quantified for purposes of modeling (e.g., frequency counts, embeddings, cosine similarity).

3. Training and evaluating models focused on predicting RIASEC descriptiveness ratings (criteria) as a function of input data (e.g., occupation descriptions, titles, task statements, importance ratings for KSAs) using the 2013-2018 Dataset formulated in Step 1.

4. Training and evaluating *ensembles* of the prediction models above to see if we could improve the prediction of interest ratings. Ensembles are simply composites of different models' predicted values. Our goal here was to identify the best-performing individual models and combine their predictions together to improve the prediction of each RIASEC dimension. We examined two stages of ensembles using the 2013-2018 Dataset formulated in Step 1. Inputs for the first-stage ensembles for a given RIASEC dimension were predictions made from our initial models for that dimension, and inputs for the second-stage ensembles for a given RIASEC dimension were predictions from the best-performing first-stage ensembles for all RIASEC dimensions.

Creating Inputs/Features for Prediction Models

The 2008-2013 Dataset included information necessary to construct the features (a machine-learning term for model inputs or predictor variables) that we included in various prediction models we examined. As we note later, not all features were included in all prediction models; here, we simply introduce how key features were constructed. Some features were already quantified in the dataset (i.e., importance ratings for KSAs and GWAs), but some of the fields in

the dataset were text and required preprocessing to quantify as numeric features (i.e., occupation descriptions, titles, task statements, interest items). Below, we describe the approaches we used to generate features from occupational descriptions, titles, and task text based on bag-of-words (BoW) methods, SBERT embeddings, and text-similarity metrics.

In all analyses involving task statements, we limited the task list for each occupation to only those task statements that were designated by O*NET as "core" tasks.[8] If an occupation did not include any core tasks (i.e., because the task list was new and O*NET had not yet received ratings of task importance, or none of the tasks satisfied O*NET's criteria for core tasks), we used all task statements associated with that occupation.

Bag-of-Words (BoW) Features

As a baseline approach to representing text as numeric data, we used the bag-of-words (BoW) methods from Putka et al. (2023) to compute frequency-based features. BoW methods account for the ways in which word usage relates to text characteristics but do not account for the semantic similarity of words (e.g., synonyms such as "vehicle" and "automobile") nor the context in which those words are used (e.g., "appendix" could refer to a bodily organ or an auxiliary portion of a document).

We prepared the text for BoW processing by concatenating the relevant text (i.e., O*NET-SOC title, occupation description, and task statements) from each O*NET-SOC into a single block of text, replacing contractions with their uncontracted long forms, removing all punctuation except for intra-word dashes, replacing symbols with the words they represent (e.g., replacing ampersands with "and"), replacing numerals with words (e.g., replacing "20" with "twenty" and "3rd" with "third"), and converting all text to lower case.

Next, we used TreeTagger (Schmid, 1994) to map each word to its lemma (i.e., its dictionary form) so different inflections and forms of a word can be considered equivalent in our analyses (e.g., "acquire," "acquiring," and "acquired" were all matched to the lemma "acquire"). After lemmatizing the text, we computed four different quantitative representations of the tokens ("token" is an NLP term for a word-like text element):

- Raw term frequencies (raw counts of the number of times each token occurred in each piece of text)
- Relative term frequencies (raw term frequencies divided by the number of tokens in each piece of text)

We used relative term frequencies in our BoW models. Raw term frequencies were simply interim values that supported the computation of relative term frequencies; we did not use those values as features in our modeling work.

SBERT Embeddings

BoW methods allow researchers to produce simple quantitative representations of text using only frequencies of word usage within documents (and across documents, in the case of TF-IDF weights). However, these methods only facilitate literal evaluations of the text; as noted earlier,

[8] See O*NET Online's Scales, Ratings, and Standardized Scores page for a description of how core tasks are defined in O*NET: https://www.onetonline.org/help/online/scales

they do not account for the semantic similarity among lemmas or the contexts in which words are used. Advances in language modeling have made it possible to overcome these limitations of BoW methods and quantify the semantic content of sentences or paragraphs in more nuanced ways. For this research, we used the Sentence BERT (SBERT) model from the "SentenceTransformers" Python library (Reimers & Gurevych, 2019); specifically, we used the "nli-distilroberta-base-v2" SBERT model. SBERT is a model for converting pieces of sentence- or paragraph-like text to scores on a set of 768 quantitative dimensions, and it was built using bidirectional encoder representations from transformers (BERT; Devlin et al., 2019) networks. HumRRO has had success using SBERT with text from the O*NET database in the past, most notably in the development of an updated related occupations framework (Dahlke et al., 2022).

We used SBERT to generate embeddings (in their original scaling used by the transformer model) for the following configurations of text[9]:

- O*NET-SOC Titles: Embedding generated for each title as a phrase-like piece of text.
- Occupation descriptions: Embeddings generated by treating each description as an intact block of text.
  - During exploratory analyses, we found this yielded better predictions than generating a separate embedding for each sentence from the descriptions and then aggregating/averaging the sentence-level embeddings within each occupation.
- Task statements: Embeddings generated by treating each task as a separate document, then aggregating/averaging the embeddings across tasks within an occupation.
  - During exploratory analyses, we found this yielded better predictions than concatenating tasks from each occupation into a paragraph-like block of text and generating embeddings for each collection of concatenated text.
- Combined occupational text (O*NET-SOC title, occupation description, and task statements): Embeddings generated for concatenated paragraph-like blocks of text consisting of occupation titles, descriptions, and tasks.

As we note later, these different configurations correspond to different feature sets we included in our prediction models.

*Cosines Between SBERT Embeddings for O*NET-SOCs and Interests*

We also used SBERT to generate embeddings for O*NET Interest Profiler Short Form RIASEC items and CABIN items (with O*NET's illustrative activities for basic interests added to the item lists of their respective CABIN dimensions) as both (a) individual items and (b) a concatenated block of item text for each dimension. We used those embeddings to create features that express the similarity between O*NET-SOCs' text and sets of text that exemplify each of the six RIASEC dimensions. We used two types of text to create features that represent RIASEC similarity: (a) O*NET text from occupations that are either very high or very low on a given RIASEC dimension and (b) item text from the Interest Profiler Short Form RIASEC measure.

[9] Later in our research effort, we also generated embeddings for other text that became available in later versions of the O*NET database after O*NET 18.0 (e.g., alternate titles, detailed work activities, intermediate work activities). Under Step 6, we explain how we used these additional types of text to further refine models developed under Step 2.

We also used features summarizing the similarity between O*NET-SOCs' text and basic interest dimensions from CABIN. We discuss each of these in the subsections that follow.

Cosines Between SBERT Embeddings for O*NET-SOCs and Exemplar Occupations

For each RIASEC dimension, we identified the O*NET-SOCs with importance ratings in the top and bottom 5% of that dimension's distribution within the subset of occupations we used to train the prediction models (see the Sample Splitting section below). Then, for each of those 12 sets of exemplar O*NET-SOCs, we computed the average cosine between the SBERT embeddings for each O*NET-SOC's text and the exemplar O*NET-SOCs' text. For these cosines, we used embeddings that represented concatenated text from O*NET-SOC titles, occupation descriptions, and task statements. We used the following formula to compute the cosine between embeddings from each pair of O*NET-SOCs:

$$Cosine_{X,Y} = \frac{\sum_{i=1}^{768}\left(SBERT_{X_i} \times SBERT_{Y_i}\right)}{\sqrt{\sum_{i=1}^{768} SBERT_{X_i}^2}\sqrt{\sum_{i=1}^{768} SBERT_{Y_i}^2}}$$

We used the average cosine between each O*NET-SOC's text and each of the 12 sets of exemplar O*NET-SOCs as features in our modeling work.

Cosines Between SBERT Embeddings for O*NET-SOCs and Interest Items

For each dimension of the RIASEC and CABIN measures (including the illustrative activities O*NET developed for each of the CABIN dimensions), we computed the cosine between the SBERT embeddings for each O*NET-SOC and each interest dimension. For these cosines, we used O*NET-SOC embeddings that represented concatenated text from O*NET-SOC titles, occupation descriptions, and task statements. On the interest side, we used embeddings we constructed by (a) generating item-level embeddings for each item on each measure and (b) averaging the item-level embeddings within each interest dimension, resulting in one aggregate SBERT embedding per interest dimension. After comparing the O*NET-SOC embeddings and the interest dimension embeddings, each O*NET-SOC had a profile of cosines that represented the similarity of that occupation's text with the text representing each of the RIASEC and CABIN dimensions.

Initial Models

We developed a set of models to predict ratings for each of the six RIASEC dimensions. Below, we first describe the models we considered and their features. We then describe how we trained and cross-validated the models. Lastly, we close this section with an evaluation of our initial models.

Specifications for Initial Models

We fit a variety of models to predict RIASEC ratings; the types of features used in these models are defined in Table 2.1. Separate models were trained to predict each RIASEC dimension. For each model, we evaluated the performance of different potential regression methods for fitting that model. We evaluated least squares (OLS) regression, sparse partial least squares (SPLS) regression, and elastic net (EN) regression for all models except those where the number of features was equal to or greater than the size of the training sample (i.e., M1-M5); in those cases, we used only SPLS and EN regression. SPLS and EN are well-suited to models in

which there are a large number of features (and even models in which the number of features exceeds the number of cases), and both allow uninformative features to be dropped from a model (we describe these methods below).

Table 2.1. Summary of Models Trained for Each RIASEC Dimension

Model	Description	Regression Methods Evaluated	# of Features	Feature Type
M1	Bag of Words (BoW) Features	SPLS, EN	1082	Occupational Titles, Descriptions, Tasks
M2	SBERT Embeddings for Concatenated Text from Titles, Descriptions, and Tasks	SPLS, EN	768	Occupational Titles, Descriptions, Tasks
M3	Title SBERT Embeddings	SPLS, EN	768	Occupational Titles
M4	Description SBERT Embeddings	SPLS, EN	768	Occupational Descriptions
M5	Task SBERT Embeddings	SPLS, EN	768	Tasks
M6	SBERT High RIASEC Occupation Similarity	OLS, SPLS, EN	6	High RIASEC Occupational Similarity
M7	SBERT Low RIASEC Occupation Similarity	OLS, SPLS, EN	6	Low RIASEC Occupational Similarity
M8	SBERT Combined High/Low RIASEC Occupation Similarity	OLS, SPLS, EN	12	High and Low RIASEC Occupational Similarity
M9	SBERT RIASEC Interest Profiler Similarity	OLS, SPLS, EN	6	IP Item Similarity
M10	SBERT CABIN Similarity	SPLS, EN	41	CABIN Item Similarity
M11	O*NET Knowledge Importance Ratings	OLS, SPLS, EN	33	Knowledge Importance
M12	O*NET Skill Importance Ratings	OLS, SPLS, EN	35	Skill Importance
M13	O*NET Ability Importance Ratings	OLS, SPLS, EN	52	Ability Importance
M14	O*NET GWA Importance Ratings	OLS, SPLS, EN	41	GWA Importance

Note. OLS = Ordinary least squares regression. SPLS = Sparse partial least squares regression. EN = Elastic net regression. # of Features = Number of features initially input into the models. Two of the regression methods we examined (SPLS and EN) perform variable selection, so the number of features in the final fitted model may be less than the starting number of features initially input into the model.

SPLS regression (Chun & Keleş, 2010) is an approach for modeling high-dimensional data (i.e., data with a large number of features), especially when those data exhibit multicollinearity. It is an improvement on partial least squares (PLS), which is a method for reducing the dimensionality of a feature set while predicting an outcome. Whereas PLS reduces the dimensionality of features within a sample without accounting for the cross-validation of that process, SPLS uses hyperparameters to regularize the dimension reduction process and limit the extent to which the model capitalizes on idiosyncrasies of a single sample. We varied the *K* hyperparameter (which determines the number of components extracted from the feature set to use as predictors in the model) from 1 to 10 in increments of 1, except when a model had fewer than 10 features (in those cases, we set the upper limit of *K* to the number of features minus 1).

We also varied the eta threshold hyperparameter from 0.0 to 0.9 in increments of 0.1 (this hyperparameter can take on values between 0 and 1).

EN regression (Friedman et al., 2010) is a regularized regression procedure that helps to avoid overfitting due to multicollinearity. It combines two other regularized regression approaches—ridge regression and LASSO (least absolute shrinkage and selection operator) regression—into a single framework that accounts for the regularization penalties from both methods. Elastic net regression uses a hyperparameter called alpha to blend the L1 (LASSO) and L2 (ridge) penalties into a single regularization penalty. The alpha hyperparameter can range from 0 (all weight to ridge) to 1 (all weight to LASSO), and its ability to mix the two penalties—or allow either ridge or LASSO to serve as a special case penalty term when alpha is 0 or 1—is what gives elastic net regression its "elasticity." We varied the alpha mixing hyperparameter from 0 to 1 in increments of 0.1, and we varied the lambda hyperparameters (which determine the severity of the regularization penalty) from 10^{-4} to 20. To constrain the number of lambda values while still covering this whole range, we defined the candidate values by raising 10 to the power of 100 equally spaced exponent values between -4 and 1 and also tested lambda values ranging from 11 to 20 in increments of 1. EN regression is sensitive to the scaling of features, and all features must be on the same scale for the method to work properly, so we standardized all features using means and *SD*s from the training sample (described in the sample splitting section below) before conducting our analyses.

After fitting models using each of the relevant regression methods for a given RIASEC dimension, we compared the cross-validated performance of each combination of model formulation and regression method. We selected the model-method combinations that exhibited the largest cross-validated correlation with interest ratings from trained human raters for use in subsequent ensemble modeling efforts. Before we describe the subsequent ensemble modeling efforts, though, we first describe the sample splitting, hyperparameter tuning, fitting, and cross-validation strategy we used to evaluate the models in Table 2.1.

Sample Splitting

To avoid overfitting our models (and subsequent ensembles), we took steps to ensure we did not capitalize on idiosyncrasies in the data during all steps of our modeling procedure. Specifically, we first split the 2008-2013 Dataset into two subsets: 75% of occupations were assigned to our "training" data set that we used to fit models, and the other 25% were assigned to our "test" data set that we used to cross-validate the performance of our models. We stratified our data splitting process by job family to ensure a comparable representation of all job families in both partitions of our data.[10] As we explored the performance of our models, this independent split of our data allowed us to use the test set to evaluate the performance of models developed using the training set and determine how the models perform when applied to unfamiliar data.

Hyperparameter Tuning

We divided our training data into five "folds" to support the development of models that involve hyperparameters. Some types of models—such as ordinary least squares (OLS) regression and logistic regression—involve estimating slope and intercept parameters based on relations in a dataset, and an analyst has no influence over those parameter estimates aside from their choice of cases to include in the dataset on which their model is trained. In other words, in these

[10] Job families are synonymous with the "Major Group" level included and described in the Standard Occupational Classification. These job families are also imbedded in the O*NET-SOC 2019 Taxonomy.

types of models, the parameters are fully determined by the data on which a model is trained; once the data have been chosen, the parameter estimates are invariant. The procedures described in this section are not relevant to these types of models.

For other types of models, such as SPLS and EN regression, the correspondence between data and parameter estimates is not as straightforward because there is not just one way in which the model can learn. These models require an analyst to select hyperparameters that govern *how* the model learns from the training data; once the data have been chosen, the parameter estimates may vary as a function of one's choice of hyperparameters. Depending on the machine learning approach, hyperparameters can define things such as the speed with which the model learns (e.g., when developing models involving decision trees, a model will learn faster when one uses fewer trees but may not perform as well as a model involving more trees) or the way in which model error/loss is quantified (e.g., ridge regression and LASSO regression require the analyst to define a λ regularization hyperparameter that defines the severity of the penalty a model incurs for having large coefficients; this penalty is added to the model's total error term that also includes residual/unexplained variance in the observations).

Whereas model parameters (e.g., regression coefficients) are learned directly from a training data set, hyperparameters must be selected by an analyst, usually through a process of experimentation known as hyperparameter tuning. There are infinite possible sets of hyperparameters one could test, so an analyst will need to decide up-front on a strategy for identifying candidate values for each hyperparameter. Hyperparameter tuning commonly involves (a) dividing the training data into *k* equal folds (where the number of folds, *k*, is chosen by the analyst) and (b) deciding how to sample sets of hyperparameters to test, typically through either a "grid search" (i.e., systematically test all possible combinations of the candidate hyperparameter values; this is the strategy we use in this research) or a "random search" (i.e., test randomly sampled combinations of the candidate hyperparameter values). In this research, we used five folds of training data and a grid search strategy. Table 2.2 shows the breakdown of our data between the training and test data sets, as well as how the training data were divided into folds.

Table 2.2. Percentage of Data Included in the Training Data, Training Folds, and Test Data

Data Segment	Training Data	Test Data
Complete Segment	75%	25%
Fold 1	15%	---
Fold 2	15%	---
Fold 3	15%	---
Fold 4	15%	---
Fold 5	15%	---

Note. All percentages express the sizes of data segments relative to the complete dataset.

After applying the data partitioning strategy from Table 2.2 to the 974 O*NET-SOCs in the 2008-2013 Dataset and stratifying our sampling approach by job family, we arrived at the analysis sample summarized in Table 2.3. The first fold of our training data segment ended up slightly larger than the other folds, but not considerably so, and our procedure for balancing job families had its intended effect of ensuring that all data segments were similar in their makeup.

Table 2.3. Sample Sizes for Job Families Across 2008-2013 Dataset Segments

Job Family		Training Data Folds						Test Data	Total
		All	1	2	3	4	5		
11	Management	44	9	9	9	8	9	15	59
13	Business and Financial Operations	38	8	8	7	8	7	13	51
15	Computer and Mathematical	25	5	5	5	5	5	8	33
17	Architecture and Engineering	54	11	10	11	11	11	17	71
19	Life, Physical, and Social Science	45	9	9	9	9	9	15	60
21	Community and Social Service	11	2	2	2	2	3	3	14
23	Legal	6	2	1	1	1	1	2	8
25	Educational Instruction and Library	46	9	9	9	10	9	15	61
27	Arts, Design, Entertainment, Sports, and Media	32	6	7	6	7	6	11	43
29	Healthcare Practitioners and Technical	64	13	13	13	13	12	22	86
31	Healthcare Support	14	3	3	3	2	3	4	18
33	Protective Service	22	5	5	4	4	4	7	29
35	Food Preparation and Serving Related	13	3	2	3	3	2	4	17
37	Building and Grounds Cleaning and Maintenance	6	1	1	2	1	1	2	8
39	Personal Care and Service	24	5	4	5	5	5	8	32
41	Sales and Related	18	4	3	4	4	3	6	24
43	Office and Administrative Support	48	10	10	10	9	9	15	63
45	Farming, Fishing, and Forestry	12	3	2	2	2	3	5	17
47	Construction and Extraction	45	9	9	9	9	9	16	61
49	Installation, Maintenance, and Repair	41	8	8	8	9	8	13	54
51	Production	84	17	16	17	17	17	28	112
53	Transportation and Material Moving	39	8	8	8	7	8	14	53
Total		731	150	144	147	146	144	243	974

Note. The number preceding the job family is the first two digits of the O*NET-SOC 2019 code corresponding to that job family.

The hyperparameter tuning process is iterative, and the total number of iterations is the product of the number of folds multiplied by the number of parameter combinations to be tested. Each instance of the process involves selecting the set of candidate hyperparameters to test, selecting which fold to set aside from one's training data as a holdout fold, training a model using the selected hyperparameters with the non-holdout training data, computing predictions on the holdout fold using the model, and evaluating the fit of the predicted values to the actual values within the holdout fold. In our case, we evaluated the fit of predictions by first computing the root mean squared error (*RMSE*) between predicted values and actual outcomes (i.e., RIASEC ratings) within each holdout fold, and then computing the mean and standard deviation of those *RMSE*s across folds. We selected the hyperparameter values that yielded the lowest mean *RMSE* (and, in the event of a tie, the smallest standard deviation of hyperparameters' *RMSE*s) for use in subsequent models. We also saved the predicted values for all holdout folds for use in our subsequent ensemble modeling efforts, as this helped us to maintain independence between the data used to train models and those used to evaluate the models.

Cross-Validation and Final Initial Model Fitting

After identifying the best-performing set of hyperparameters (if applicable to the modeling procedure), we trained a model using the complete training data set and cross-validated that model using the test data set. We then fit a final model on the complete dataset and saved this final model for potential future use. We also saved the predicted values for the test data for use in evaluating the performance of our ensemble models.

Evaluation of Initial Models

As noted above, we first experimented with OLS, SPLS, and EN regression when fitting our initial set of models. We applied all three methods to each feature set, except when the number of features was equal to or greater than the size of the training sample (i.e., M1-M5), as those methods can accommodate analyses involving more features than observations. The best-performing methods and hyperparameter values for those methods for each model are summarized in Table C.1 in Appendix C. Except for Model 1, the EN method consistently outperformed the SPLS method.

For the best-performing specification (method) for each model, we calculated *RMSE* and multiple *R* metrics, and we have organized the results for these metrics in Tables 2.4 and 2.5, respectively. In all our modeling efforts, we used *RMSE* as the primary fit metric for comparing competing machine learning methods, hyperparameter combinations, and feature sets and selecting our final model for each RIASEC dimension. We also report multiple *R* values as it is a standardized metric familiar to researchers (analogous to a criterion-related validity estimate) and can facilitate comparison to other research and benchmarks (e.g., comparison to interrater reliability estimates). We did not use the multiple *R* statistics to make decisions during our modeling workflow, as we encountered a situation early in our model development process that suggested focusing on maximizing *R* can have undesirable consequences with some types of models. Specifically, we noticed an analysis in which we obtained a large multiple *R* result after tuning our hyperparameters, but although the linear relationship was strong in the standardized sense, when we plotted the observed and predicted values, the predicted values did not have the same scaling as the observed values. After identifying that issue, we adopted *RMSE* as our primary fit metric because it characterizes the strength of the association between two sets of values while also being sensitive to differences in the scaling of those values. After altering our models to focus on minimizing *RMSE*, we did not encounter any additional problems with predictions having incorrect scales.

Table 2.4. Cross-Validated RMSE Results for Best Specifications for Initial 14 Models for Each RIASEC Dimension

Model		*N*		*# of*	Average Cross-Validated *RMSE* Across Hold-Out Training Folds							Cross-Validated *RMSE* in Testing Data						
		Train	Test	Features	*M*	R	I	A	S	E	C	*M*	R	I	A	S	E	C
M5	Task SBERT Embeddings	731	243	768	.948	.932	1.134	.756	.823	1.024	1.019	.947	.898	1.099	.782	.835	1.025	1.041
M2	SBERT Embeddings for Concatenated Text from Titles, Descriptions, and Tasks	731	243	768	.936	.913	1.086	.799	.831	.959	1.027	.949	.898	1.074	.750	.896	1.001	1.073
M8	SBERT Combined High/Low RIASEC Occupation Similarity	731	243	12	.974	1.012	1.131	.817	.807	1.059	1.018	.975	1.012	1.165	.817	.815	1.006	1.037
M10	SBERT CABIN Similarity	731	243	41	.978	.962	1.161	.825	.835	1.065	1.022	.979	.949	1.166	.799	.804	1.070	1.083
M4	Description SBERT Embeddings	731	243	768	1.011	1.025	1.167	.860	.871	1.084	1.058	.998	.989	1.100	.809	.932	1.085	1.071
M6	SBERT High RIASEC Occupation Similarity	731	243	6	1.016	1.066	1.188	.823	.886	1.090	1.041	1.010	1.067	1.203	.816	.901	1.016	1.058
M1	Bag of Words (BoW) Features	731	243	1082	.981	.962	1.183	.840	.853	1.014	1.032	1.031	.926	1.208	.900	.927	1.049	1.178
M11	O*NET Knowledge Importance Ratings	641	219	33	1.103	1.033	1.249	.916	1.085	1.192	1.145	1.054	1.079	1.157	.784	.888	1.292	1.123
M14	O*NET GWA Importance Ratings	641	219	41	1.105	.990	1.312	.951	1.051	1.186	1.138	1.110	.987	1.298	.941	1.016	1.265	1.155
M9	SBERT RIASEC Interest Profiler Similarity	731	243	6	1.132	1.111	1.357	.935	.996	1.335	1.060	1.149	1.110	1.391	.979	.990	1.335	1.090
M7	SBERT Low RIASEC Occupation Similarity	731	243	6	1.149	1.046	1.354	.909	1.094	1.349	1.141	1.182	1.018	1.495	.890	1.181	1.357	1.153
M12	O*NET Skill Importance Ratings	628	211	35	1.196	1.121	1.177	1.221	1.146	1.226	1.282	1.186	1.151	1.143	1.134	1.114	1.314	1.258
M3	Title SBERT Embeddings	731	243	768	1.141	1.211	1.254	.967	1.073	1.128	1.213	1.191	1.269	1.301	1.008	1.121	1.249	1.196
M13	O*NET Ability Importance Ratings	641	219	52	1.201	.992	1.411	.995	1.217	1.374	1.219	1.204	1.031	1.372	.968	1.271	1.399	1.180

Note. # of Features = Number of features initially input into the model. *M* = Average cross-validated *RMSE* across RIASEC dimensions. Models are sorted in ascending order of mean cross-validated *RMSE* in testing data. *RMSE* values are shaded along a green-red color gradient to facilitate interpretation (lower values—indicating better model performance—are shaded green; higher values—indicating poorer model performance—are shaded red).

Table 2.5. Cross-Validated Multiple R Results for Best Specifications for Initial 14 Models for Each RIASEC Dimension

Model		*N*		# of	Average Cross-Validated *R* Across Hold-Out Training Folds							Cross-Validated *R* in Testing Data						
		Train	Test	Features	*M*	R	I	A	S	E	C	*M*	R	I	A	S	E	C
M5	Task SBERT Embeddings	731	243	768	.84	.89	.81	.85	.91	.83	.74	.85	.91	.84	.85	.91	.84	.74
M2	SBERT Embeddings for Concatenated Text from Titles, Descriptions, and Tasks	731	243	768	.84	.90	.83	.83	.91	.85	.73	.85	.91	.85	.86	.90	.86	.72
M8	SBERT Combined High/Low RIASEC Occupation Similarity	731	243	12	.83	.87	.81	.82	.92	.81	.74	.84	.88	.82	.83	.92	.85	.75
M10	SBERT CABIN Similarity	731	243	41	.83	.89	.80	.82	.91	.81	.73	.84	.90	.81	.84	.92	.83	.72
M4	Description SBERT Embeddings	731	243	768	.82	.87	.80	.80	.90	.81	.72	.84	.89	.84	.84	.89	.83	.72
M6	SBERT High RIASEC Occupation Similarity	731	243	6	.82	.86	.79	.82	.90	.80	.72	.83	.87	.80	.83	.90	.85	.74
M1	Bag of Words (BoW) Features	731	243	1082	.83	.89	.79	.81	.91	.83	.73	.82	.90	.80	.79	.89	.84	.68
M11	O*NET Knowledge Importance Ratings	641	219	33	.78	.87	.75	.79	.84	.74	.66	.82	.88	.81	.86	.91	.74	.70
M14	O*NET GWA Importance Ratings	641	219	41	.77	.88	.72	.77	.85	.75	.66	.79	.90	.76	.78	.87	.75	.68
M9	SBERT RIASEC Interest Profiler Similarity	731	243	6	.76	.85	.71	.76	.87	.68	.71	.77	.86	.72	.75	.88	.72	.71
M7	SBERT Low RIASEC Occupation Similarity	731	243	6	.75	.87	.71	.77	.84	.67	.65	.76	.88	.67	.80	.82	.71	.67
M3	Title SBERT Embeddings	731	243	768	.76	.81	.76	.74	.85	.79	.59	.76	.81	.77	.73	.84	.76	.64
M12	O*NET Skill Importance Ratings	628	211	35	.72	.85	.78	.59	.82	.73	.53	.75	.85	.81	.67	.84	.72	.61
M13	O*NET Ability Importance Ratings	641	219	52	.72	.88	.67	.74	.80	.64	.60	.75	.88	.72	.77	.78	.68	.66

Note. # of Features = Number of features initially input into the model. *M* = Average cross-validated *R* across RIASEC dimensions. Models are sorted in ascending order of mean cross-validated *R* in testing data. *R* values are shaded along a green-red color gradient to facilitate interpretation (higher values—indicating better model performance—are shaded green; lower values—indicating poorer model performance—are shaded red).

In general, the level of prediction obtained by these models was strong and in line with existing benchmarks. For example, Table 2.6 compares cross-validated test set multiple *R*'s for what was the best-performing model on average across RIASEC dimensions (M5: Task SBERT Embeddings) to three benchmarks:

- Weighted average single-rater reliability (ICC[C,1]) for sample raters in the 2008-2013 O*NET interest data collections. This approximates the expected correlation one would expect to see between any two trained raters selected at random.
- Weighted average interrater reliability (ICC[C,3]) for sample raters in the 2008-2013 O*NET interest data collections. This approximates the expected correlation one would expect to see between mean ratings provided by two randomly selected groups of three trained raters—effectively, the operational reliability of the published interest ratings in O*NET. This reliability estimate reflects the theoretical upper bound on the correlation between predicted and observed interest ratings.
- Putka et al. (2023) – The cross-validated multiple *R* for the BoW-based model reported in the Putka et al. article that modeled O*NET interest ratings.

Table 2.6. Comparison of Initial Model 5 Performance to Existing Benchmarks

Benchmark/Model	*M*		R	I	A	S	E	C
Weighted Average Single Rater Reliability ICC(C,1)	.79		.86	.79	.78	.86	.78	.64
Weighted Average Interrater Reliability ICC(C,3)	.91		.95	.92	.91	.95	.92	.84
Putka et al., (2023) Model Cross-Validated *R*	.84		.90	.83	.83	.92	.85	.73
Model 5: Task SBERT Embeddings Cross-Validated Test Set *R*	.85		.91	.84	.85	.91	.84	.74

Note. *M* = Average across RIASEC dimensions. ICC(C,1) and ICC(C,3) statistics reflect interrater reliability estimates for interest ratings in the 2008-2013 Dataset.

Model 5's cross-validated test set multiple *R*'s consistently exceeded the correlation one might expect to see between any two trained raters and approached the upper bound on validity implied by ICC(C,3). The levels of test set cross-validity observed for Model 5 were also comparable to those reported by Putka et al. (2023)

Though the initial models showed promising results across the RIASEC dimensions, the predictions for Conventional interests tended to have the weakest correspondence to their target values. We designed our modeling workflow around the potential for ensembles to make up for the deficiencies of individual initial models, so we proceeded to develop first-stage ensembles in which the predictions from our initial models were treated as features.

First-Stage Ensembles

The modeling process described in the sections above was only an initial step in our prediction strategy. After training the models and determining which machine learning methods performed the best for predicting each RIASEC dimension, we experimented with ways to combine the predictions from sets of models into ensemble predictions to see if we could improve prediction. Table 2.7 shows the various sets of models we combined into composites using ensemble models for each RIASEC dimension. For each ensemble, we evaluated combining model

predictions using OLS, SPLS, and EN regression to predict interest ratings to see which one produced better results.

Table 2.7. Summary of Ensembles Trained for Each RIASEC Dimension

	Base Models (Titles, Descriptions, and Tasks)				Separate High/Low RIASEC Occupation Models		Combined High/Low RIASEC Occupation Model	Interest Profiler Item Similarity Model	CABIN Item Similarity Model	KSA Importance Model			GWA Importance Model
Ensemble	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
E1	X	X	X	X									
E2	X	X	X	X	X	X							
E3	X	X	X	X			X						
E4	X	X	X	X	X	X		X					
E5	X	X	X	X			X	X					
E6	X	X	X	X	X	X			X				
E7	X	X	X	X			X		X				
E8	X	X	X	X	X	X		X	X				
E9	X	X	X	X			X	X	X				
E10	X	X	X	X						X	X	X	X
E11	X	X	X	X	X	X				X	X	X	X
E12	X	X	X	X			X			X	X	X	X
E13	X	X	X	X	X	X		X		X	X	X	X
E14	X	X	X	X			X	X		X	X	X	X
E15	X	X	X	X	X	X			X	X	X	X	X
E16	X	X	X	X			X		X	X	X	X	X
E17	X	X	X	X	X	X		X	X	X	X	X	X
E18	X	X	X	X			X	X	X	X	X	X	X

First-Stage Ensemble Training and Cross-Validation Process

We used the same training and test data as we used in our initial modeling analyses and the same folds of data as we used in our five-fold hyperparameter tuning process (for SPLS and EN regression). To maintain independence between our training and test samples in both our hyperparameter tuning and our cross-validation analyses, we saved the predictions we generated for each set of four training folds during our modeling process to use as features in our ensemble models. As we set aside each fold of our training data to use as a holdout sample during hyperparameter tuning, we used the predictions generated from models trained on the other four folds as features in our ensembling process. For example, when we built an ensemble model that omitted the first fold of training data, we used predictions from models trained on folds 2–5 to both train the model and evaluate the model using data from fold 1.

Evaluation of First-Stage Ensembles

We experimented with OLS, SPLS, and EN methods to develop first-stage ensemble models that used predictions from our initial models as features. The best-performing methods and hyperparameter values for each ensemble are summarized in Table C.2 in Appendix C, and the corresponding RMSE and multiple *R* fit metrics are summarized in Tables 2.8 and 2.9, respectively.

Table 2.8. Cross-Validated RMSE Results for Best Specifications for 18 First-Stage Ensembles for Each RIASEC Dimension

Ensemble	*N*		# of Features	Average Cross-Validated *RMSE* Across Hold-Out Training Folds							Cross-Validated *RMSE* in Testing Data						
	Train	Test		*M*	R	I	A	S	E	C	*M*	R	I	A	S	E	C
16	628	211	10	.841	.777	1.010	.706	.753	.855	.947	.845	.824	.944	.663	.759	.921	.956
12	628	211	9	.842	.780	1.005	.706	.756	.855	.952	.845	.824	.942	.660	.768	.921	.954
18	628	211	11	.843	.778	1.011	.710	.754	.854	.948	.847	.820	.955	.671	.758	.922	.956
11	628	211	10	.845	.779	1.011	.710	.759	.856	.956	.847	.817	.940	.669	.773	.925	.958
15	628	211	11	.843	.777	1.011	.710	.755	.855	.952	.848	.816	.952	.670	.762	.925	.960
17	628	211	12	.843	.777	1.011	.709	.755	.854	.953	.848	.816	.952	.671	.761	.925	.960
14	628	211	10	.843	.781	1.011	.707	.756	.855	.950	.849	.821	.956	.674	.766	.921	.953
13	628	211	11	.845	.780	1.012	.709	.759	.855	.955	.849	.817	.954	.671	.770	.926	.957
10	628	211	8	.846	.779	1.007	.710	.764	.857	.961	.850	.824	.941	.663	.780	.927	.965
5	731	243	6	.875	.860	1.039	.723	.756	.899	.973	.879	.846	1.028	.702	.792	.923	.981
3	731	243	5	.876	.860	1.039	.723	.755	.899	.978	.881	.846	1.028	.702	.805	.923	.982
9	731	243	7	.875	.860	1.041	.723	.756	.899	.970	.882	.846	1.032	.702	.781	.946	.987
7	731	243	6	.874	.860	1.041	.723	.751	.899	.971	.883	.846	1.032	.702	.787	.945	.985
8	731	243	8	.876	.860	1.043	.723	.759	.900	.973	.887	.846	1.035	.702	.803	.949	.986
6	731	243	7	.877	.860	1.043	.723	.759	.900	.975	.888	.846	1.035	.702	.806	.948	.990
4	731	243	7	.879	.860	1.047	.723	.763	.903	.977	.888	.846	1.034	.702	.815	.951	.981
2	731	243	6	.879	.860	1.046	.723	.762	.902	.982	.890	.846	1.034	.702	.819	.950	.989
1	731	243	4	.882	.859	1.049	.722	.771	.903	.987	.895	.845	1.027	.700	.829	.954	1.017

Note. # of Features = Number of features initially input into the ensemble. *M* = Average cross-validated *RMSE* across RIASEC dimensions. Ensembles are sorted in ascending order of mean cross-validated *RMSE* in testing data. *RMSE* values are shaded along a green-red color gradient to facilitate interpretation (lower values—indicating better ensemble performance—are shaded green; higher values—indicating poorer ensemble performance—are shaded red).

Table 2.9. Cross-Validated Multiple R Results for Best Specifications for 18 First-Stage Ensembles for Each RIASEC Dimension

Ensemble	*N*		# of Features	Average Cross-Validated *R* Across Hold-Out Training Folds							Cross-Validated *R* in Testing Data						
	Train	Test		*M*	R	I	A	S	E	C	*M*	R	I	A	S	E	C
12	628	211	9	.87	.93	.85	.88	.93	.88	.78	.89	.93	.88	.90	.93	.88	.80
16	628	211	10	.87	.93	.84	.88	.93	.88	.78	.88	.93	.87	.90	.93	.88	.80
14	628	211	10	.87	.93	.84	.88	.93	.88	.78	.88	.93	.87	.90	.93	.88	.80
18	628	211	11	.87	.93	.84	.88	.93	.88	.78	.88	.93	.87	.90	.93	.88	.80
11	628	211	10	.87	.93	.84	.88	.93	.88	.78	.88	.93	.88	.90	.93	.88	.80
13	628	211	11	.87	.93	.84	.88	.93	.88	.78	.88	.93	.87	.90	.93	.87	.80
17	628	211	12	.87	.93	.84	.88	.93	.88	.78	.88	.93	.87	.90	.93	.87	.80
15	628	211	11	.87	.93	.84	.88	.93	.88	.78	.88	.93	.87	.90	.93	.87	.80
10	628	211	8	.87	.93	.84	.88	.93	.88	.77	.88	.93	.88	.90	.93	.87	.80
5	731	243	6	.86	.91	.84	.86	.93	.87	.76	.87	.92	.86	.88	.92	.88	.77
3	731	243	5	.86	.91	.84	.86	.93	.87	.76	.87	.92	.86	.88	.92	.88	.77
9	731	243	7	.86	.91	.84	.86	.93	.87	.77	.87	.92	.86	.88	.93	.87	.77
7	731	243	6	.86	.91	.84	.86	.93	.87	.77	.87	.92	.86	.88	.92	.87	.77
4	731	243	7	.86	.91	.84	.86	.93	.87	.76	.87	.92	.86	.88	.92	.87	.78
8	731	243	8	.86	.91	.84	.86	.93	.87	.76	.87	.92	.86	.88	.92	.87	.77
6	731	243	7	.86	.91	.84	.86	.93	.87	.76	.87	.92	.86	.88	.92	.87	.77
2	731	243	6	.86	.91	.84	.86	.93	.87	.76	.87	.92	.86	.88	.92	.87	.77
1	731	243	4	.86	.91	.84	.86	.93	.87	.76	.87	.92	.86	.88	.92	.87	.76

Note. # of Features = Number of features initially input into the ensemble. *M* = Average cross-validated *R* across RIASEC dimensions. Ensembles are sorted in ascending order of mean cross-validated *R* in testing data. *R* values are shaded along a green-red color gradient to facilitate interpretation (higher values—indicating better ensemble performance—are shaded green; lower values—indicating poorer ensemble performance—are shaded red).

Although the first-stage ensembles did improve prediction relative to the best-performing initial models (particularly the ensembles that included KSA/GWA models, i.e., E10-E18), the gains were not that great. We hypothesize the lack of substantial gains may be a reflection of high correlations among predictions from the initial models. Therefore, we computed correlations among model predictions and summarized the distributions of those correlations in Table 2.10.

Table 2.10. Summary of Correlations Among Predictions from Initial Models Used as Features in First-Stage Ensembles

Model Range	$N_{Occupations}$	$k_{Correlations}$	Statistic	R	I	A	S	E	C
M2 – M10 (Text Models Only)	974	36	Mean	.93	.88	.89	.93	.86	.87
			SD	.04	.06	.06	.04	.07	.07
			Min	.85	.76	.76	.84	.70	.73
			Max	.99	.98	1.00	.99	.99	.98
M2 – M14 (Text Models and KSA/GWA Rating Models)	839	78	Mean	.91	.82	.82	.89	.80	.78
			SD	.04	.08	.09	.05	.08	.11
			Min	.83	.64	.61	.78	.65	.55
			Max	.99	.98	1.00	.99	.99	.98

We found the average correlation between initial model predicted values ranged from .86 for Enterprising interests to .93 for Realistic and Social interests across models contributing to the text-only ensembles (i.e., E1-E9) and ranged from .78 for Conventional interests to .91 for Realistic interests across models contributing to the text+KSA/GWA ensembles (i.e., E10-E18).[11] Ensembles gain the most predictive advantage over input models when predictions from input models are less correlated with one another.

Furthermore, we noticed that even with ensembling, the multiple *R* values for Conventional interests reported in Table 2.9 continued to be lower than those for the other five RIASEC dimensions. The ensembles were not as effective as we had hoped at making up for the deficiencies of the initial models for Conventional interests. Our analysis strategy included two additional opportunities to improve the prediction of Conventional interest ratings (as well as the ratings for other dimensions) through our second-stage ensemble models and an additional phase of modeling work (the latter is described in the Step 6 section of this report).

To gain more insight into how the best-performing first-stage ensemble models were functioning and how they were assigning weight to their features, we retrained all initial models and the best-performing first-stage ensembles on the full set of occupations in the 2008-2013 Dataset. We then gathered the model coefficients from these fully trained models and performed general dominance analyses to determine how each initial model contributed to the ensemble predictions (i.e., in terms of the proportion of the ensembles' explained variance that was attributable to each initial model). These analyses did not require cross-validation, as they focus

[11] For purposes of this analysis, correlations were based on retrained versions of all initial models on the full set of occupations in the 2008-2013 Dataset (i.e., the predictions used for the regression coefficient and general dominance results described below).

on the internal functioning of the models and do not involve evaluations of the models' ability to generalize to new samples.

The best-performing first-stage ensembles used for these analyses (based on average cross-validated RMSE across holdout training folds reported in Table 2.8) as well as for input into the second-stage ensembles we discuss next were as follows[12]:

For O*NET-SOCs without KSA/GWA data:

- Realistic: Ensemble 1
- Investigative: Ensemble 3
- Artistic: Ensemble 1
- Social: Ensemble 7
- Enterprising: Ensemble 3
- Investigative: Ensemble 9

For O*NET-SOCs with KSA/GWA data:

- Realistic: Ensemble 15
- Investigative: Ensemble 12
- Artistic: Ensemble 12
- Social: Ensemble 7
- Enterprising: Ensemble 17
- Investigative: Ensemble 16

Tables 2.11 and 2.12 show the coefficients and general dominance weights for best-performing first-stage ensembles for O*NET-SOCs without KSA/GWA data (Table 2.11) and for O*NET-SOCs with KSA/GWA data (Table 2.12). Regardless of the ensemble, the general dominance weights varied little across initial models that contributed to that ensemble, indicating that the initial models tend to contribute similarly to the ensemble predictions. We hypothesize that this again is due to high correlations among predictions from the initial models, effectively resulting in the weight assigned to input models being distributed more evenly.

[12] Note, we differentiate between two types of best-performing ensembles here because only the text-only ensembles (i.e., E1-E9) could be used for O*NET-SOCs without KSA/GWA data, whereas any ensemble (E1-E18) could be applied to O*NET-SOCs with KSA/GWA data.

*Table 2.11. Regression Coefficients and Relative Importance Estimates for Best First-Stage Ensembles for Each RIASEC Dimension for O*NET-SOCs without KSA/GWA Data*

Ensemble Feature	Realistic		Investigative		Artistic		Social		Enterprising		Conventional	
	B	*RI*	*B*	*RI*	*B*	*RI*	*B*	*RI*	*B*	*RI*	*B*	*RI*
Intercept	-.456	---	-.584	---	-.332	---	-.237	---	-.596	---	-.898	---
Model 2	.408	.266	.235	.211	.342	.267	.184	.171	.235	.213	.186	.148
Model 3	.117	.219	.242	.187	.123	.210	.189	.154	.241	.189	.202	.121
Model 4	.126	.248	.240	.198	.237	.252	.185	.166	.241	.199	.161	.134
Model 5	.444	.267	.235	.206	.451	.271	.174	.170	.235	.206	.181	.156
Model 6	---	---	---	---	---	---	---	---	---	---	---	---
Model 7	---	---	---	---	---	---	---	---	---	---	---	---
Model 8	---	---	.217	.198	---	---	.175	.170	.219	.194	.161	.153
Model 9	---	---	---	---	---	---	---	---	---	---	.164	.140
Model 10	---	---	---	---	---	---	.177	.169	---	---	.161	.148

Note. *B* = Regression coefficient. *RI* = Relative importance reflecting the proportion of the ensemble R^2 attributable to the given model based on a general dominance analysis. "---" indicates that predictions from the given model were not included in the best-performing ensemble for the given RIASEC dimension.

*Table 2.12. Regression Coefficients and Relative Importance Estimates for Best First-Stage Ensembles for Each RIASEC Dimension for O*NET-SOCs with KSA/GWA Data*

Ensemble Feature	Realistic		Investigative		Artistic		Social		Enterprising		Conventional	
	B	*RI*	*B*	*RI*	*B*	*RI*	*B*	*RI*	*B*	*RI*	*B*	*RI*
Intercept	-.475	---	-.873	---	-.546	---	-.237	---	-.670	---	-1.500	---
Model 2	.207	.104	.144	.128	.157	.131	.184	.171	.189	.113	.147	.117
Model 3	.127	.087	.147	.107	.168	.112	.189	.154	.227	.103	.160	.094
Model 4	.166	.099	.149	.117	.162	.124	.185	.166	.182	.103	.130	.108
Model 5	.192	.098	.145	.126	.155	.129	.174	.170	.167	.101	.142	.122
Model 6	-.037	.085	---	---	---	---	---	---	.044	.087	---	---
Model 7	-.017	.087	---	---	---	---	---	---	-.037	.060	---	---
Model 8	---	---	.136	.118	.150	.122	.175	.170	---	---	.127	.114
Model 9	---	---	---	---	---	---	---	---	-.003	.060	---	---
Model 10	.084	.093	---	---	---	---	.177	.169	.075	.092	.127	.114
Model 11	.103	.087	.142	.107	.154	.112	---	---	.095	.074	.137	.097
Model 12	.045	.080	.132	.110	.000	.061	---	---	.078	.071	.138	.069
Model 13	.133	.089	.131	.087	.156	.102	---	---	.100	.058	.127	.077
Model 14	.099	.091	.141	.100	.154	.107	---	---	.079	.077	.126	.089

Note. *B* = Regression coefficient. *RI* = Relative importance reflecting the proportion of the ensemble R^2 attributable to the given model based on a general dominance analysis. "---" indicates that predictions from the given model were not included in the best-performing ensemble for the given RIASEC dimension.

Second-Stage Ensembles

The final set of models we developed during this step of our research were a second stage of ensembles that used the predictions from the best-performing first-stage ensembles (identified earlier) as features for predicting each RIASEC dimension. Predictions from these best-performing first-stage ensembles for each RIASEC dimension served as inputs into our second-stage ensembles. We used the same approach to training and cross-validation as we used for our first-stage ensembles (i.e., the predictions we use as features in these models will preserve the independence of the training and test samples). In this case, however, we formed the ensembles using OLS regression only since the second-stage ensemble for each RIASEC dimension consisted of only six predictors (i.e., the best first-stage ensemble prediction for the target RIASEC dimension, and the best first-stage ensemble predictions for the other five RIASEC dimensions). This approach is meant to capitalize on the circumplex structure of the RIASEC interest model in which the dimensions are not orthogonal; adjacent dimensions tend to be more positively correlated, while more distal dimensions tend to be weakly or negatively correlated. In previous research, we have found that ensemble models involving features based on all six RIASEC dimensions can increment the prediction of individual RIASEC dimensions (Dahlke & Putka, 2022).

Before training the second-stage ensembles, we retrained the initial models and best-bet first-stage ensembles on the entire training data set to generate the required features for the second-stage ensembles. Beyond the fact that they use the same features to predict all RIASEC dimensions' ratings, the second-stage ensembles also differ from our other models in that we only trained them using OLS regression, and, therefore, they do not require hyperparameter tuning. Even so, we continued using the same five-fold cross-validation procedure as we used with the initial models and first-stage ensembles to (a) maintain consistency in how we produced evaluative metrics and (b) provide preliminary cross-validated fit metrics based on the training sample before applying models to the test sample.

We have summarized the RMSE and multiple *R* fit metrics for the second-stage ensembles in Tables 2.13 and 2.14, respectively. To facilitate a comparison of the performance of the best-performing first-stage ensembles and previously introduced benchmarks, Table 2.15 provides a side-by-side comparison of these ensembles and benchmarks. Examination of Table 2.15 reveals that the second-stage ensembles appear to offer no clear advantage over the first-stage ensembles in terms of prediction regardless of dimension (in contrast to Dahlke & Putka, 2022).

Lastly, we summarized the coefficients and general dominance weights for all second-stage ensembles in Table 2.16. These general dominance analysis results reveal that using the circumplex structure to form ensembles had the intended effect of parsing variance across RIASEC dimensions in each of the dimension-specific models. Regardless of whether KSA/GWA features were available for inclusion in the models, about 50% of the variance in second-stage Realistic interest predictions was attributable to first-stage Realistic interest predictions, about 83%-85% of the variance in second-stage Investigative interest predictions was attributable to first-stage Investigative interest predictions, and, for the remaining dimensions, about two-thirds of the variance in second-stage predictions was attributable to the first-stage predictions for the target interest dimension.

Table 2.13. Cross-Validated RMSE Results for Second-Stage Ensembles

Ensemble	*N*		# of Features	Average Cross-Validated *RMSE* Across Hold-Out Training Folds							Cross-Validated *RMSE* in Testing Data						
	Train	Test		*M*	R	I	A	S	E	C	*M*	R	I	A	S	E	C
Text Only	731	243	6	.876	.867	1.042	.724	.751	.899	.972	.879	.851	1.022	.700	.788	.925	.987
Text and KSA/GWA	628	211	6	.839	.778	1.001	.701	.763	.847	.946	.845	.807	.947	.649	.793	.919	.954

Table 2.14. Cross-Validated Multiple R Results for Second-Stage Ensembles

Ensemble	*N*		# of Features	Average Cross-Validated *R* Across Hold-Out Training Folds							Cross-Validated *R* in Testing Data						
	Train	Test		*M*	R	I	A	S	E	C	*M*	R	I	A	S	E	C
Text Only	731	243	6	.86	.91	.84	.86	.93	.87	.76	.87	.92	.86	.88	.92	.88	.77
Text and KSA/GWA	628	211	6	.87	.93	.85	.88	.93	.88	.78	.88	.93	.87	.91	.92	.88	.80

Table 2.15. Comparison of Best First-Stage and Second-Stage Ensembles to Existing Benchmarks

Benchmark/Model	*M*	R	I	A	S	E	C
Weighted Average Single Rater Reliability ICC(C,1)	.79	.86	.79	.78	.86	.78	.64
Weighted Average Interrater Reliability ICC(C,3)	.92	.95	.92	.91	.95	.92	.84
Putka et al (2023) Model Cross-Validated *R*	.84	.90	.83	.83	.92	.85	.73
Best-Performing Text First-Stage Ensemble Cross-Validated Test Set *R*	.87	.92	.86	.88	.93	.88	.77
Best-Performing Text + KSA/GWA First-Stage Ensemble Cross-Validated Test Set *R*	.89	.93	.88	.90	.93	.87	.80
Text Only Second-Stage Ensemble Cross-Validated Test Set *R*	.87	.92	.86	.88	.92	.88	.77
Text + KSA/GWA Second-Stage Ensemble Cross-Validated Test Set *R*	.88	.93	.87	.91	.92	.88	.80

Note. *R* values are shaded along a green-red color gradient to facilitate interpretation (higher values—indicating better performance—are shaded green; lower values—indicating poorer performance—are shaded red).

Table 2.16. Regression Coefficients and Relative Importance Estimates for Second-Stage Ensembles

Ensemble	Input	Realistic		Investigative		Artistic		Social		Enterprising		Conventional	
		B	*RI*	*B*	*RI*	*B*	*RI*	*B*	*RI*	*B*	*RI*	*B*	*RI*
Text Only	Intercept	-3.023	---	-1.262	---	-.695	---	-.828	---	-.711	---	-.860	---
Text Only	Realistic (E1)	1.209	.525	.090	.030	.059	.104	.076	.207	.088	.169	.055	.057
Text Only	Investigative (E3)	.070	.015	1.029	.836	.003	.014	.007	.007	.033	.056	.018	.027
Text Only	Artistic (E1)	.137	.091	.067	.016	1.058	.680	.022	.047	.008	.013	.024	.148
Text Only	Social (E7)	.150	.166	.028	.009	.003	.043	1.044	.639	.023	.027	.072	.061
Text Only	Enterprising (E3)	.084	.158	.034	.076	.025	.015	.043	.031	1.087	.670	-.025	.063
Text Only	Conventional (E9)	.187	.045	.095	.033	.045	.145	.029	.070	-.049	.065	1.088	.643
Text and KSA/GWA	Intercept	-3.331	---	-.209	---	-.880	---	-.887	---	-1.663	---	-2.627	---
Text and KSA/GWA	Realistic (E15)	1.254	.493	.041	.034	.088	.110	.082	.216	.157	.168	.192	.066
Text and KSA/GWA	Investigative (E12)	.086	.016	1.005	.849	-.003	.017	.011	.007	.035	.038	.043	.017
Text and KSA/GWA	Artistic (E12)	.143	.104	-.005	.018	1.085	.658	.018	.052	.050	.016	.119	.153
Text and KSA/GWA	Social (E7)	.156	.169	.013	.009	.007	.044	1.052	.622	.029	.034	.158	.051
Text and KSA/GWA	Enterprising (E17)	.084	.166	-.026	.064	.007	.016	.049	.044	1.171	.682	.004	.055
Text and KSA/GWA	Conventional (E16)	.188	.052	.015	.026	.057	.155	.026	.060	.002	.061	1.202	.657

Note. Input = Intercept and best-performing first-stage ensemble for each RIASEC dimension (noted in parentheses). *B* = Regression coefficient. *RI* = Relative importance reflecting the proportion of the ensemble R^2 attributable to the given model based on a general dominance analysis. "---" indicates that predictions from the given model were not included in the best-performing ensemble for the given RIASEC dimension. *RI* values are shaded along a green-red color gradient to facilitate interpretation (higher values are shaded green; lower values are shaded green).

Final Consolidation of Step 2 Modeling Results

After training all models and ensembles planned for this first phase of analyses, we compared the performance of all first- and second-stage ensemble models for each RIASEC dimension to identify the best-performing ensemble for (a) occupations that had only text-based features available and (b) occupations that had both text-based features and KSA/GWA ratings available. Up to this point in our modeling process, all evaluations had been based on average cross-validated performance across folds within the training data to maintain independence between model comparisons made during the model development process (i.e., for determining which machine learning methods, hyperparameters, and—for ensembles—feature sets to use; and to select the final, best-forming models from this phase of predictive analyses). In this final consolidation, we used fit metrics for the 25% of holdout test data to decide which models to carry forward as empirical best-bet benchmarks.

Table 2.17 shows the RMSE and multiple *R* results for each RIASEC dimension's best-performing ensemble by feature availability. We based our model-selection evaluations on unbounded predicted values without imposing the 1–7 range that defines the rating scale for O*NET's RIASEC importance ratings. We do, however, show the metrics for predicted ratings after constraining them to the 1–7 range to provide an indication of the impact of that transformation on the quality of predictions. We based model selection on unbounded predictions because the "final" models identified during this phase of predictive modeling are, in fact, only preliminary prediction models within the complete scope of our research. The truly final models were identified later in Step 6 when both predictive power and practical considerations were factored into the final model selection.

Table 2.17. Cross-Validity Estimates for Best Performing Ensembles

Feature Availability	Dimension	Best Ensemble	Unbounded Predictions		Predictions Bounded Between 1 and 7	
			RMSE	*R*	*RMSE*	*R*
Text Only	Realistic	1st Stage E1	.845	.92	.825	.92
Text Only	Investigative	2nd Stage E	1.022	.86	1.011	.87
Text Only	Artistic	1st Stage E1	.700	.88	.678	.89
Text Only	Social	1st Stage E9	.781	.93	.772	.93
Text Only	Enterprising	1st Stage E3	.923	.88	.913	.88
Text Only	Conventional	1st Stage E4	.981	.78	.968	.78
Text and KSA/GWA	Realistic	2nd Stage E	.807	.93	.789	.93
Text and KSA/GWA	Investigative	1st Stage E11	.940	.88	.935	.88
Text and KSA/GWA	Artistic	2nd Stage E	.649	.91	.623	.92
Text and KSA/GWA	Social	1st Stage E18	.758	.93	.751	.93
Text and KSA/GWA	Enterprising	2nd Stage E	.919	.88	.907	.88
Text and KSA/GWA	Conventional	1st Stage 14	.953	.80	.935	.81

Step 3: Generating Preliminary OIPs and High-Point Codes

Upon the conclusion of the model development and evaluation process above, we used the best-performing ensemble for each RIASEC dimension (identified in Table 2.17) to generate predicted RIASEC ratings (OIPs) for the 923 data-level occupations in the O*NET 27.1 Database (i.e., the O*NET database that was most current as of this step in our development process, see O*NET Database Release Archives). Based on these ratings, we then assigned up to three high-point codes for each occupation using the following steps originally developed and defined by Rounds et al. (1999):

1. Convert RIASEC ratings to proportions within each occupation.
2. Assign initial high-point codes for each occupation, such that the RIASEC dimension with the highest proportion was assigned the 1st high-point position for the occupation, the RIASEC dimension with the 2nd highest proportion was assigned the 2nd high-point position for the occupation, and the RIASEC dimension with the 3rd highest proportion was assigned the 3rd high-point position for the occupation.[13]
3. Retain only those high points for an occupation where the RIASEC dimension assigned to that high point had a proportion greater than .17 (i.e., a variable high-point code system). For example, if a third high-point code for an occupation listed a RIASEC dimension with a rating proportion of .15, no third high-point code was listed for that occupation.

Note that OIPs and high-point codes assigned at this stage in the process were preliminary and were designed to reflect empirically optimal values based on the initial modeling work done in Step 2. We used these values to help identify occupations to target for sampling as part of the analyst and expert data collection conducted as part of this work (see Step 5). Specifically, we used these ratings to identify occupations where we had less certainty about the quality of prediction either because (a) the O*NET occupation had no published RIASEC rating in O*NET 27.1 or (b) predictions made by the best-performing ensemble were discrepant from the RIASEC ratings of record in the 2008-2013 Dataset. Our process for identifying these occupations is discussed next in Step 4.

[13] Note, that Rounds et al. (1999) describe rules used for resolving ties among ratings for the top three RIASEC dimensions for purposes of assigning high-point codes. In the case of our predicted ratings, there were no ties for the top three RIASEC dimensions for any occupations, so no tie breaking rules were needed.

Step 4: Identifying Occupations for Inclusion in Analyst-Expert Rating Data Collections

As part of the current effort, we aimed to gather O*NET analyst and RIASEC expert ratings for a subset of the 923 data-level O*NET-SOCs to aid in further evaluation and refinement the prediction models developed based on the 2008-2013 Dataset. Our intent here was not to gather human ratings for all 923 data-level occupations but rather strategically target a subset of occupations where we had less certainty about the quality of the predicted ratings. A total of 269 O*NET-SOC occupations were identified for inclusion in the data collection. Table 4.1 provides a summary of inclusion criteria and the number of occupations that met each criterion. Tables 4.2 and 4.3 provide a comparison of the representativeness of the sample of 269 occupations relative to the full set of 923 data-level occupations in terms of job zone and job family representation.[14]

The inclusion criteria shown in Table 4.1 are not mutually exclusive. They were implemented sequentially and reflect the research team's aim to incrementally build up a set of occupations for human review. As Table 4.1 reveals, most of the criteria we implemented reflected different potential standards for agreement among machine-based (predicted) and human ratings that we considered to fill out the list of occupations (see Criteria 3 – 10).

Table 4.1. Inclusion Criteria for O*NET-SOC Data Level Occupations in Analyst/Expert Data Collection

Inclusion Criterion	# of Occupations Included for this Criterion	Cumulative # of Occupations Included
1. Data-level occupation in O*NET 27.1 does not appear in the O*NET-SOC 2019 to O*NET-SOC 2010 crosswalk.	4	4
2. Data-level occupation in O*NET 27.1 that does not have interest data in O*NET 27.1.	45	49
3. Data-level occupation in O*NET 27.1 with a 1-to-1 crosswalk match to an O*NET-SOC 2010 occupation, AND that has more than two RIASEC dimensions that have predictions that fall outside the 95% standard error of measurement (SEM) based confidence interval around the published interest ratings.	60	109
4. Data-level occupation in O*NET 27.1 with a 1-to-1 crosswalk match to an O*NET-SOC 2010 occupation, AND that has two RIASEC dimensions that have predictions that fall outside the 95% standard error of measurement (SEM) based confidence interval around the published interest ratings, AND that has (a predicted-observed profile correlation < .80 AND a predicted-observed 1st high-point code that does not match).	12	120
5. Data-level occupation in O*NET 27.1 with a 1-to-1 crosswalk match to an O*NET-SOC 2010 occupation, AND that has two RIASEC dimensions that have predictions that fall outside the 95% standard error of measurement (SEM) based confidence interval around the published interest ratings, AND that has (a predicted-observed profile correlation < .80 OR a predicted-observed 1st high-point code that does not match).	48	156

[14] O*NET Job Zones group occupations into one of five categories based on levels of education, experience, and training necessary to perform the occupation (Rivkin & Craven, 2021).

Table 4.1. (Continued)

Inclusion Criterion	# of Occupations Included for this Criterion	Cumulative # of Occupations Included
6. Data-level occupation in O*NET 27.1 with a 1-to-1 crosswalk match to an O*NET-SOC 2010 occupation, AND that has two or more RIASEC dimensions that have predictions that fall outside the 99% standard error of measurement (SEM) based confidence interval around the published interest ratings.	77	178
7. Data-level occupation in O*NET 27.1 with a 1-to-1 crosswalk match to an O*NET-SOC 2010 occupation, AND that has (a predicted-observed profile correlation < .80 AND a predicted-observed 1st high-point code that does not match).	25	180
8. Data-level occupation in O*NET 27.1 with a 1-to-1 crosswalk match to an O*NET-SOC 2010 occupation, AND that has a predicted-observed profile correlation < .80.	42	182
9. Data-level occupation in O*NET 27.1 with a 1-to-1 crosswalk match to an O*NET-SOC 2010 occupation, AND that has one RIASEC dimension that has predictions that fall outside the 99% standard error of measurement (SEM) based confidence interval around the published interest ratings, AND a predicted-observed 1st high-point code that does not match.	47	205
10. Data-level occupation in O*NET 27.1 with more than one crosswalk match to an O*NET-SOC 2010 occupation, AND that has more than one RIASEC dimension that has predictions that fall outside the 95% standard error of measurement (SEM) based confidence interval around the published interest ratings.	64	269

Note. Inclusion criteria are not mutually exclusive and were implemented sequentially. The Cumulative # of Occupations Included column reflects the cumulative number of occupations selected for inclusion upon implementing the given criterion.

Comparison of occupations for inclusion in the analyst-expert rating data collections relative to the full set of 923 data-level O*NET-SOCs revealed good coverage of all O*NET job zones and job families. For example, Table 4.2 shows that occupations selected for inclusion in the data collection were generally comparable in terms of their distribution across job zones relative to the full set of 923 occupations, with slightly less representation of job zone two occupations, and slightly more representation of job zone three and four occupations. Similarly, Table 4.3 shows that occupations selected for inclusion in the data collection were generally comparable in terms of their distribution across job families, with slightly less representation of Production occupations and slightly more representation of Healthcare Practitioners and Technical occupations.

*Table 4.2. Representativeness of Occupations Selected for Inclusion in the Analyst/Expert Data Collection with Respect to O*NET Job Zone*

Job Zone	All O*NET 27.1 Data-Level Occupations			O*NET 27.1 Data Level Occupations Selected for Inclusion in 2023 Data Collection			
	n	%		*n*	%		Δ%
1: Little or no preparation needed	32	3.5		6	2.2		-1.2
2: Some preparation needed	289	31.3		53	19.7		-11.6
3: Medium preparation needed	220	23.8		76	28.3		4.4
4: Considerable preparation needed	224	24.3		86	32.0		7.7
5: Extensive preparation needed	158	17.1		48	17.8		0.7
Total	923	100.0		269	100.0		0.0

Note. Δ% = % of all O*NET 27.1 data-level occupations in the given job zone minus % of O*NET 27.1 data-level occupations selected for inclusion in the 2023 data collection.

Table 4.3. Representativeness of Occupations Selected for Inclusion in the Analyst/Expert Data Collection with Respect to Job Family

Job Family	All O*NET 27.1 Data-Level Occupations			O*NET 27.1 Data Level Occupations Selected for Inclusion in 2023 SME Data Collection			
	n	%		*n*	%		Δ%
Architecture and Engineering	56	6.1		14	5.2		-0.9
Arts, Design, Entertainment, Sports, and Media	40	4.3		19	7.1		2.7
Building and Grounds Cleaning and Maintenance	8	0.9		1	0.4		-0.5
Business and Financial Operations	48	5.2		22	8.2		3.0
Community and Social Service	14	1.5		2	0.7		-0.8
Computer and Mathematical	36	3.9		20	7.4		3.5
Construction and Extraction	61	6.6		8	3.0		-3.6
Educational Instruction and Library	62	6.7		22	8.2		1.5
Farming, Fishing, and Forestry	12	1.3		4	1.5		0.2
Food Preparation and Serving Related	16	1.7		7	2.6		0.9
Healthcare Practitioners and Technical	89	9.6		36	13.4		3.7
Healthcare Support	19	2.1		3	1.1		-0.9
Installation, Maintenance, and Repair	50	5.4		6	2.2		-3.2
Legal	7	0.8		3	1.1		0.4
Life, Physical, and Social Science	60	6.5		18	6.7		0.2
Management	56	6.1		21	7.8		1.7
Office and Administrative Support	51	5.5		6	2.2		-3.3
Personal Care and Service	31	3.4		13	4.8		1.5
Production	107	11.6		11	4.1		-7.5
Protective Service	26	2.8		12	4.5		1.6
Sales and Related	22	2.4		5	1.9		-0.5
Transportation and Material Moving	52	5.6		16	5.9		0.3
Total	923	100.0		269	100.0		0.0

Note. Δ% = % of all O*NET 27.1 data-level occupations in the given job zone minus % of O*NET 27.1 data-level occupations selected for inclusion in the 2023 data collection in the given job zone.

Step 5: Collecting and Evaluating O*NET Analyst and Expert RIASEC Ratings

HumRRO recruited sets of trained O*NET analysts, as well as academics with expertise in the study of vocational interests to provide RIASEC ratings for the 269 occupations identified in Step 4. The reason for obtaining ratings from separate groups of analysts and experts was motivated by a need to (a) evaluate whether using the existing cadre of O*NET analysts may provide a reliable and valid source of vocational interest ratings should the need arise in the future, and (b) obtain expert ratings to serve as criteria not only for evaluating O*NET analyst ratings but also in subsequent steps of our RIASEC modeling process that focused on evaluating the inclusion of new features that became available in O*NET post-2013 (i.e., features that were not available in the 2008-2023 Dataset used for modeling in Step 2, namely alternate titles, intermediate work activities, detailed work activities).

Overview of Rater Recruitment and Training

We recruited six O*NET analysts with several years of experience in the O*NET Data Collection Program who have provided skill and ability ratings for O*NET. We aimed to recruit analysts who (a) provided high-quality ratings in past data collection efforts, (b) came from a diverse range of demographic backgrounds, and (c) were timely in their provision of past ratings. Additionally, we recruited three academic experts who professionally study and publish on the topic of vocational interests to serve as our expert raters. Those individuals were Dr. James Rounds of the University of Illinois Urbana-Champaign's Department of Psychology (our primary consultant on this effort), Dr. Rong Su of the University of Iowa's Tippie College of Business, and Dr. Kevin Hoff of Michigan State University's Department of Psychology.

Rater Training

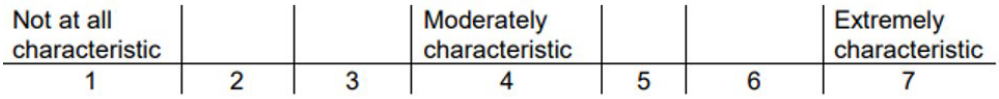
Rater training involved the following activities: (a) providing an overview of the rating task, (b) describing Holland's RIASEC model and how RIASEC interests are defined within O*NET, (c) introducing the scale on which RIASEC dimensions are rated in O*NET, (c) engaging in a RIASEC familiarization exercise, and (d) rating and discussing an initial subset of occupations for rater calibration purposes. In advance of training, HumRRO shared all training materials with Dr. Rounds for review, editing, and approval. HumRRO provided separate training for both groups of raters. For both groups of raters, the training process was comparable in terms of the substantive nature of training materials presented, with the exception of the length of instructional time for experts, which was reduced due to their substantial familiarity with RIASEC.

During training, we provided raters with an overview of Holland's RIASEC model and subsequently reviewed each RIASEC dimension in-depth. More specifically, HumRRO provided updated O*NET RIASEC definitions, keywords associated with a particular RIASEC dimension, and illustrative activities and occupations for each dimension (Rounds et al., 2023).

To provide context for the rating task, we introduced raters to the two types of occupation-level interest data in O*NET: (a) Occupational Interest Profiles (OIPs) and (b) Interest high-point codes. Raters were informed the ratings they were making for OIPs are used to generate interest high-point codes for occupations. Since OIPs are dependent on the accuracy of raters, it was emphasized that reaching agreement was very important. HumRRO then provided an illustrative example of how RIASEC ratings for a given occupation result in OIPs, and relatedly, high-point codes.

Raters were informed they would be using the same interest ratings scale as the one originally used by Rounds et al. (1999) that asks, “How descriptive and characteristic is the given Holland work environment of this occupation?” where a rating of 1 is “Not at all characteristic,” and a rating of 7 is “Extremely characteristic” (see Figure 5.1).

Figure 5.1. O*NET RIASEC Dimension Rating Scale



Following the introduction of RIASEC and the different types of interest data in O*NET, we engaged in a RIASEC familiarization exercise with raters that attempted to get raters thinking about the relationship between RIASEC and occupational work environments. Appendix D provides instructions for exercise (see *RIASEC Familiarization Exercise Instructions*).

Overview of Rating Process

Upon completion of the training described above, we provided raters with materials to complete their RIASEC ratings for each occupation. These materials consisted of two Excel-based files: (a) an occupational information booklet and (b) a master ratings booklet. The occupational information booklet consisted of occupations along with their descriptions and tasks sorted by job family.[15] Occupation descriptions and tasks were drawn from O*NET 27.1 – the latest available O*NET database at the time of the rater training. The tasks included for the majority of occupations (n = 221) reflected those considered “core” to the occupation by O*NET (i.e., relevance ≥ 67% and a mean importance rating ≥ 3.0) and were sorted in order of importance to that occupation. A limited number of occupations (n = 48) did not have relevance or importance ratings associated with their tasks in O*NET 27.1, so for those occupations, all tasks associated with that occupation were presented in alphabetical order. Raters were instructed to review all occupational information before making a RIASEC rating. Appendix D provides the instructions given to raters for making their ratings and an example of the rating sheet where they entered their ratings.

The rating process consisted of three phases:

(1) Initial calibration and group discussion: Raters first independently rated a subset of 10 occupations. As noted above, this rating was done as part of the initial analyst and expert training session. The subset of occupations selected for this phase was designed to represent a diverse range of occupations and job families with respect to their standing on the RIASEC dimensions. After raters made their initial ratings for these occupations, we discussed the ratings as a group and discussed areas of disagreement to develop a clearer shared understanding of the RIASEC categories. Dr. Rounds participated in the calibration discussion with analysts and offered perspective when raters shared why they made certain occupation-RIASEC associations. Raters were permitted to independently update their initial ratings for the 10 occupations following the group discussion.

[15] The rationale behind sorting occupations by job family was to reduce cognitive load and help raters better see differences across occupations both between and within a job family.

(2) **Follow-up calibration and group discussion**: Following the initial calibration above, raters were assigned and independently rated another subset of occupations asynchronously. In this phase, analysts rated 50 more occupations, and experts rated 15 more occupations. Once again, the subset of occupations selected for this phase was designed to represent a diverse range of occupations and job families with respect to their standing on the RIASEC dimensions. Following the raters' submission of their ratings for this second subset of occupations, we conducted analyses to identify areas of disagreement. We first reviewed ratings from the experts, who were in unanimous agreement on the top three high-point codes for over half of the occupations they rated. HumRRO shared the agreement results with the experts and concluded that a second re-calibration meeting with them would not be needed. Next, we reviewed ratings from the analysts, who exhibited greater evidence of disagreement than the experts. Thus, we followed up with a group discussion session with the analyst to review their ratings.

In preparation for the group discussion with analysts, we flagged occupations where there was greater disagreement among analysts on their RIASEC ratings and prioritized those for discussion. We also attended to any themes or patterns in ratings, especially those that seemed counterintuitive. For example, Conventional ratings tended to be consistently elevated across occupations and job families. When we subsequently discussed this observation with analysts, they shared the observation that a wide array of O*NET occupations' core tasks concern activities such as recording, documenting, or reporting. More specifically, one rater argued that Social Workers would be a highly Conventional occupation considering the number of core Social Worker tasks that involve reporting, documenting, and maintaining client records. Additionally, raters indicated that post-secondary occupations, like "Atmospheric, Earth, Marine, and Space Sciences Teachers, Postsecondary" involve core task statements that rarely mention hard sciences and primarily focus on administrative aspects of the job like grading, maintaining student records, etc. As such, analysts tended to rate the Conventional dimension highly for several more socially oriented occupations. In light of this rating tendency, we sought guidance from Dr. Rounds, who provided the following advice on the matter: at lower grade levels, teaching is a social activity. As teachers move into higher grade levels and the objective of teaching is with respect to a school subject, the high-point codes generally begin with the educational subject taught (e.g., if the subject being taught is an art-related subject, the first high-point may be Artistic, whereas if the subject being taught is a science-related subject the high-point codes might begin with Investigative). As such, raters should consider the level at which a subject is being taught in making ratings. Additionally, although core task statements might include conventional-oriented activities, it is important to focus on the job descriptive information.

HumRRO began the group discussion session by asking analysts to share their mental processes for making ratings, any gains in efficiencies learned, and challenges encountered while making ratings. Following that discussion, raters were re-trained on a reduced version of the original training slides, which paid particular focus on a RAISEC construct review and the input from Dr. Rounds, provided above. Following the group discussion, analysts were permitted to independently update their initial ratings for the occupations.

(3) **Independent rating of remaining occupations**: Upon completion of the calibration phases above, raters were assigned and independently rated the remaining occupations (out of the full set of 269) they had yet to rate (209 for analysts, 244 for experts). All final ratings from analysts and experts were submitted to HumRRO by the end of May 2023.

Rating Data Review and Cleaning

Once all final ratings were obtained from both groups of raters, we made several checks to ensure the information captured in individual rating workbooks was both accurate and complete. This involved ensuring each occupation-RIASEC had a rating and following up with individuals when incomplete. These checks also involved ensuring ratings were within the expected scale range (i.e., 1-7) for every occupation-RIASEC combination. We developed a master rating file with ratings from each group of raters.

Given our plans to use the expert ratings as a standard for evaluating the analyst ratings, as well as for refining our RIASEC prediction models later in the project (see Step 6), we conducted an additional layer of screening and refinement on the expert ratings. In line with the procedure and rationale employed to refine the RIASEC ratings in the 2008-2013 Dataset in Step 1, we used the three rules developed in Step 1 for flagging the 1,614 occupation-RIASEC combinations (269 occupations x 6 RIASEC dimensions) where disagreement among the three experts who provided the ratings was considered meaningful:

- Rule 1: Range across raters was greater than or equal to four (on a one-to-seven rating scale) AND if two of the experts were less than or equal to one rating point away from each other. Twenty-one out of 1,614 (1.3%) occupation-RIASEC combinations met this rule. The occupation-RIASEC combinations that met this rule spanned 19 of the 269 (7.0%) occupations rated.
- Rule 2: Range across raters was greater than or equal to four AND if two of the raters were greater than one rating point away from each other. Fourteen out of 1,614 (0.8%) occupation-RIASEC combinations met this rule. The occupation-RIASEC combinations that met this rule spanned 14 of the 269 (5.2%) occupations rated.
- Rule 3: Range across raters was equal to three AND if two of the raters gave the same rating. Forty-seven out of 1,614 (2.9%) occupation-RIASEC combinations met this rule. The occupation-RIASEC combinations that met this rule spanned 43 of the 269 (15.9%) occupations rated.

Based on the rules above, we made the following adjustments to expert RIASEC ratings before further analyses:

- Trimmed means (or exact value, if same) were used in instances where an occupation-RIASEC combination was flagged for Rule 1, where the outlying rater's rating was removed prior to calculating the trimmed mean.
- The expert ratings from Dr. Rounds were used in instances where an occupation-RIASEC was flagged for Rule 2.
- The modal (i.e., agree upon) rating was used in instances where an occupation-RIASEC combination was flagged for Rule 3.

Note that if an occupation-RIASEC combination was not flagged based on one of the rules above, the mean rating across the three experts was used as the final rating for a given RIASEC dimension. For analysts, the mean rating across the six analysts was used as the final rating for a given RIASEC dimension.

Evaluation of Ratings

We first calculated descriptive statistics for the analyst and cleaned expert RIASEC ratings and examined standardized within-occupation differences between analysts and cleaned expert ratings. Next, we examined the reliability of RIASEC dimension ratings as well as the reliability (i.e., consistency) and absolute agreement (i.e., interchangeability) of RIASEC rating profiles furnished by analysts and experts, respectively. Lastly, we examined analyst and expert RIASEC ratings through a multitrait-multimethod correlation lens to evaluate patterns of convergence and discrimination among ratings. Except where noted below, all of the analyses above were conducted on post-calibration ratings and, for experts, post-refinement via the previously described data cleaning rules.

Basic Descriptives and Mean Differences

Table 5.1 presents descriptive summaries for each RIASEC dimension by rating source (i.e., analyst, expert). For analysts and experts, the highest means were observed for Conventional interests, and the lowest means were for Artistic interests.

Table 5.1. Descriptive Statistics for RIASEC Dimensions by Rater Type

Dimension	*M*		*SD*		5th Percentile		95th Percentile	
	Analyst	Expert	Analyst	Expert	Analyst	Expert	Analyst	Expert
Realistic	3.17	4.11	1.76	1.88	1.00	1.00	6.17	6.67
Investigative	3.29	3.60	1.52	1.81	1.17	1.00	5.50	6.00
Artistic	1.83	2.04	1.31	1.47	1.00	1.00	4.00	4.40
Social	3.96	3.29	1.25	1.62	2.00	1.13	5.67	6.00
Enterprising	3.84	3.21	1.27	1.72	2.07	1.00	5.83	6.00
Conventional	4.57	4.60	1.07	1.14	2.83	3.00	6.17	6.33

Note. *N* = 269. Ratings were made on a 7-point scale ranging from 1 (not at all characteristic) to 7 (extremely characteristic).

In addition to computing basic descriptives, we also computed within-occupation standardized mean differences between analyst and expert ratings. For any given RIASEC dimension, these standardized mean differences are calculated by first taking the difference between analyst and expert ratings for each occupation (analyst – expert), then calculating the mean difference and standard deviation of differences across occupations. The mean difference is then divided by the standard deviation of differences to arrive at the within-occupation standardized mean differences that are reported in Table 5.2.

Table 5.2 illustrates that analysts tended to rate Social interests and Enterprising interests moderately higher than experts (as indicated by positive within-occupation differences in the .70 range) and that experts tended to rate Realistic interests moderately higher than analysts (as indicated by negative within-occupation differences of -.89). Differences between analysts and experts with respect to ratings of Investigative, Artistic, and Conventional instruments appeared small to negligible.

Table 5.2. Within-Occupation Standardized Mean Differences between Rater Types

RIASEC	Analyst-Expert
Realistic	-.89
Investigative	-.29
Artistic	-.27
Social	.78
Enterprising	.67
Conventional	-.03

Note. *N* = 269.

Reliability and Agreement

Next, we examined the reliability (i.e., consistency) and absolute agreement (i.e., interchangeability) of the analyst and expert RIASEC ratings. For these analyses, we examined post-calibration ratings and, for experts, pre-refinement via the previously described data cleaning rules. Table 5.3 shows the interrater reliability and agreement estimates for each RIASEC dimension. Two pairs of reliability and agreement coefficients are provided for each rater type. ICC(C,1) reflects the estimated reliability of a single-rater's rating for the given RIASEC dimension (among raters of a given type). Effectively, ICC(C,1) is comparable to the expected correlation one would expect to see between two different raters selected at random. ICC(C,*k*) reflects the estimated reliability of the mean rating for a given RIASEC dimension (across *k* raters of a given type). Effectively, ICC(C,k) is comparable to the expected correlation one would expect to see between mean ratings provided by two randomly selected groups of *k* raters. In the case of analysts, the number of raters (*k*) equals six, and in the case of experts, the number of raters (*k*) equals three. Similarly, ICC(A,1) reflects the estimated absolute agreement of a single-rater's rating for the given RIASEC dimension (among raters of a given type). ICC(A,*k*) reflects the estimated absolute agreement of the mean rating for a given RIASEC dimension (across *k* raters of a given type).

Of prime interest in Table 5.3 are the ICC(C,6) and ICC(C,3) values for analysts and experts, respectively, as these reflect the reliability of the mean ratings – which would serve as the actual ratings for each RIASEC dimension. Focusing on these values, we see that ICC(C,6) values for analysts range from .80 (Conventional) to .92 (Artistic), and ICC(C,3) for experts ranged from .74 (Conventional) to .94 (Artistic). All of these values indicate acceptable levels of interrater reliability but are slightly lower than the levels of interrater reliability for the mean expert ratings in the 2008-2013 Dataset summarized in Step 2. Recall from Table 2.6 that for RIASEC ratings in the 2008-2013 Dataset, ICC(C,3) values for experts ranged from .84 (Conventional) to .95 (Realistic, Social). Not surprisingly, if one compares the ICC(C,1) values for analysts and expert ratings, the ICC(C,1) values for experts are notably higher. This was expected given that we would expect the correlation between RIASEC ratings of any two experienced experts selected randomly to be more highly correlated than the correlation between RIASEC ratings of any two less experienced analysts selected at random. Thus, to offset that greater level of consistency among analysts, one must sample more analysts than experts to achieve mean RIASEC ratings that are comparable in their reliability across rater types.

Table 5.3. Interrater Reliability and Agreement for RIASEC Dimensions by Rater Type

Dimension	Analyst				Expert			
	ICC(C,1)	ICC(C,6)	ICC(A,1)	ICC(A,6)	ICC(C,1)	ICC(C,3)	ICC(A,1)	ICC(A,3)
Realistic	.63	.91	.58	.89	.82	.93	.82	.93
Investigative	.61	.90	.50	.86	.78	.91	.77	.91
Artistic	.67	.92	.64	.91	.83	.94	.83	.94
Social	.50	.86	.45	.83	.77	.91	.75	.90
Enterprising	.45	.83	.43	.82	.77	.91	.77	.91
Conventional	.41	.80	.38	.79	.49	.74	.43	.70

Note. *N* = 269.

Beyond the reliability of ratings for the individual RIASEC dimensions, we also examined the reliability (i.e., consistency) and absolute agreement (i.e., interchangeability) of the within-occupation RIASEC profiles for the analyst and expert raters. In contrast to the reliability and agreement statistics for each dimension where occupations serve as the target of measurement, here our focus is on reliability and agreement for each of the 269 occupations where RIASEC dimensions serve as the target of measurement, that is, we address how consistent raters are in terms of their ordering of RIASEC dimensions for any given occupation. Table 5.4 shows the results of our analyses. Specifically, it provides descriptive statistics summarizing ICC statistics calculated for each of the 269 occupations. As with the findings above, the mean RIASEC profiles provided by both analysts and experts exhibit good levels of reliability (Analyst ICC(C,6) = .88, Expert ICC(C,3) = .91), with the 5th percentile of these ICC(C,k) values for both rater types both exceeding .70.

Table 5.4. Reliability and Agreement for RIASEC Profiles by Rater Type

Statistic	Analyst				Expert			
	ICC(C,1)	ICC(C,6)	ICC(A,1)	ICC(A,6)	ICC(C,1)	ICC(C,3)	ICC(A,1)	ICC(A,3)
Mean	.61	.88	.61	.88	.80	.91	.80	.91
SD	.17	.11	.17	.11	.13	.08	.13	.07
Min	.01	.07	.02	.08	.16	.36	.17	.38
Max	.89	.98	.89	.98	.98	.99	.98	.99
Percentiles								
5th	.30	.72	.31	.72	.56	.79	.56	.80
10th	.40	.80	.40	.80	.62	.83	.64	.84
20th	.47	.84	.48	.85	.70	.87	.71	.88
30th	.52	.87	.53	.87	.75	.89	.76	.90
40th	.58	.89	.58	.89	.79	.92	.79	.92
50th	.63	.91	.63	.91	.83	.94	.83	.93
60th	.68	.93	.68	.93	.85	.95	.85	.95
70th	.72	.94	.71	.94	.88	.96	.87	.95
80th	.77	.95	.76	.95	.90	.97	.90	.96
90th	.82	.96	.80	.96	.93	.98	.93	.97
95th	.85	.97	.83	.97	.95	.98	.94	.98

Note. *N* = 269.

Convergence among Rater Types

Lastly, we examined analyst and expert RIASEC ratings through a multitrait-multimethod correlation lens to evaluate patterns of convergence and discrimination among ratings. The purpose of doing so was to evaluate evidence of convergent and discriminant validity for the ratings (Campbell & Fiske, 1959). Table 5.5 provides correlations among analyst and expert RIASEC ratings.

Convergent validity is indicated by high correlation among the same RIASEC dimension as rated by analysts and experts (i.e., monotrait-heteromethod correlations, highlighted in yellow in Table 5.5). Table 5.5 shows strong evidence of convergence as monotrait-heteromethod correlations for all RIASEC dimensions all exceed .80 except for Conventional interests (r = .61). Note that when interpreting the analyst-expert correlation for Conventional interests, it is important to remember that this interest was the least reliably measured among both analysts and experts. When we correct this correlation for unreliability in analyst and expert ratings using the ICC(C,6) (analyst) and ICC(C,6) (expert) values for Conventional interests in Table 5.3, it increases to a corrected value of .71.

Discriminant validity and freedom from "common method" variance are indicated by relatively low correlations among different RIASEC dimensions that share a rater type in common (i.e., heterotrait-monomethod correlations, highlighted in blue in Table 5.5) and that don't clearly exceed the correlations among different RIASEC dimensions that don't share a rater type (i.e., heterotrait-heteromethod correlations, highlighted in green in Table 5.5). The pattern of results in Table 5.5 provides evidence of discriminant validity and little evidence of common method variance. Specifically, the average heterotrait-monomethod correlation among analyst ratings was -.13, the average heterotrait-monomethod correlation among expert ratings was -.11, and the average heterotrait-heteromethod correlation was -.10. The average heterotrait-monomethod correlations for both analysts and experts were also comparable to the average heterotrait-monomethod correlations calculated among the interest ratings in O*NET 27.1. Specifically, the average correlation among interest dimensions in O*NET 27.1 is -.12 (N = 874). Thus, the results here are quite comparable to what we'd expect based on the full set of occupations on which the current sample of 269 is based.

Table 5.5. Multitrait-Multimethod Correlations for RIASEC Dimensions by Rater Type

		Analyst						Expert					
		R	I	A	S	E	C	R	I	A	S	E	C
Analyst	R	1.00											
	I	-.26	1.00										
	A	-.04	-.39	1.00									
	S	-.49	-.02	.16	1.00								
	E	-.35	-.26	.00	-.02	1.00							
	C	-.49	.31	-.33	-.12	.31	1.00						
Expert	R	.84	-.06	-.09	-.35	-.44	-.47	1.00					
	I	-.37	.81	-.19	.04	-.10	.44	-.17	1.00				
	A	-.18	-.25	.85	.14	.12	-.20	-.24	-.03	1.00			
	S	-.43	.08	.14	.85	-.09	-.06	-.36	.16	.13	1.00		
	E	-.35	-.24	.02	.18	.85	.19	-.45	-.14	.10	.07	1.00	
	C	-.26	.20	-.34	-.29	.12	.61	-.25	.14	-.32	-.33	.00	1.00

Note. N = 269. Monotrait-heteromethod correlations are highlighted in yellow. Heterotrait-monomethod correlations are highlighted in blue. Heterotrait-heteromethod correlations are highlighted in green.

Step 6: Refining and Evaluating Final RIASEC Prediction Models for Future Use

Our model-building effort in Step 2 focused on understanding which models tended to work best from the perspective of maximizing the prediction of RIASEC ratings based on data available in the 2008-2013 Dataset. In this step, we shifted our focus to identifying a model for each RIASEC dimension that balanced predictive value with parsimony, then evaluated whether we could improve that model through the introduction of additional features that became available in O*NET post-2013. Specifically, the models we developed in Step 2 were limited to feature sets that were available in the O*NET Database in the 2008 to 2013 timeframe. Additional sources of features were added in later years; specifically, Detailed Work Activities (DWAs) and Intermediate Work Activities (IWAs) were introduced in the O*NET 18.1 Database (March 2014), and Alternate Titles (ATs) were introduced in the O*NET 20.1 Database (October 2015). Thus, as part of the current step, we first aimed to identify a model from Step 2 that balanced prediction with practical implementation considerations and then used that model as a baseline in this step to see if adding DWA, IWA, or AT-based features could increase its level of prediction.

Identifying a Baseline Model that Balances Prediction and Practical Considerations

Our first step was to review the modeling results from Step 2 and identify a model that provided good levels of prediction yet also satisfied practical implementation considerations. The practical factors we considered were: (a) the inputs to the model would always be available upon an occupation's introduction to O*NET (e.g., an occupation's title, description, and task statements are always available when a new occupation is introduced to O*NET, but knowledge, skill, ability, and GWA ratings are not), and (b) the model did not add additional layers of complexity unless it resulted in notable gains in prediction. Based on our review and after consultation with Dr. Rounds, we recommended the Center adopt Ensemble 1 as the baseline model of choice for each RIASEC dimension. Recall from Step 2, Ensemble 1 reflected a linear combination of predictions from four models defined by the following features (a) SBERT embeddings of an occupation's title (Model 3), (b) SBERT embeddings of an occupation's description (Model 4) (c) SBERT embeddings of an occupation's tasks (Model 5), and (d) SBERT embeddings of the concatenation of an occupation's title, description, and tasks (Model 2). The advantages we saw in Ensemble 1 included the following:

- It offered levels of prediction for each RIASEC prediction that nearly matched the level of prediction afforded by the best-performing ensembles for each RIASEC dimension identified in Step 2 (see Table 6.1),
- It was based on information that would be available upon a new occupation's introduction to O*NET (thus allowing that occupation to be nearly instantly profiled on RIASEC upon its addition to the O*NET database).
- It was simple in that it was fully based on SBERT embeddings of three simple elements of O*NET occupational text: title, description, and tasks.

Using Ensemble 1 as our new baseline, we proceeded to evaluate whether adding in new AT, DWA, and/or IWA features could improve predictions.

Table 6.1. Test Set Cross-Validated R for Ensemble 1 vs. Best Ensembles from Step 2 for Each RIASEC Dimension

Ensemble	Cross-Validated *R* in Testing Data							
	M		**R**	**I**	**A**	**S**	**E**	**C**
Ensemble 1	.87		.92	.86	.88	.92	.87	.76
Best Ensemble Step 2 - Text Only	.88		.92	.86	.88	.93	.88	.78
Best Ensemble Step 2 - Text & KSA/GWA	.89		.93	.88	.91	.93	.88	.80

Note. *M* = Average cross-validated *R* across RIASEC dimensions. Ensemble 1 cross-validated R values are drawn from Table 2.8. Best Ensemble Step 2 cross-validated R values are drawn from Table 2.15 (unbounded predictions). *R* values are shaded along a green-red color gradient to facilitate interpretation (higher values—indicating better ensemble performance—are shaded green; lower values—indicating poorer ensemble performance—are shaded red).

Developing Models that Consider AT, DWA, and IWA Features

Our first step in model development involved creating features for ATs, DWAs, and IWAs to add to Ensemble 1-based predictions. Like the text-based features in Step 2, all the features we used in this stage of our modeling analyses were SBERT embeddings we computed using the "nli-distilroberta-base-v2" model via the "SentenceTransformers" Python library (Reimers & Gurevych, 2019). We computed an SBERT embedding for each AT, DWA, and IWA from each occupation, then averaged the embeddings for each input text type within each O*NET-SOC. For each O*NET-SOC, we aggregated the SBERT embeddings for individual sentence- or phrase-like text into three vectors of average embeddings: one each for ATs, DWAs, and IWAs.

Modeling Procedure

Whereas in Step 2 of our research process, we had a total of 974 O*NET-SOCs in our analysis sample, our analysis sample for this phase of modeling only consisted of the 269 O*NET-SOCs for which new expert ratings were gathered as part of Step 5. This reduction in sample size required a different strategy for model training and cross-validation than we used in Step 2, as we did not have enough cases to (a) retrain Ensemble 1 on the new data, nor (b) support a robust cross-validation strategy based on the previous approach. If we had repeated the analysis procedures we established in Step 2 and used 75% of cases as a training sample and the remaining 25% as a testing sample, we would have had 202 occupations in our training sample and only 67 in our testing sample. Dividing a relatively small sample in this way would result in cross-validated fit estimates impacted by substantially more sampling error than in Step 2, making it much more difficult to identify the best-performing model configurations with a reasonable degree of certainty.

We made two changes to our modeling strategy in light of the more limited sample sizes referenced above. First, given that we were not in a position to retrain Ensemble 1 on the new set of data, we instead focused on whether we could use AT, DWA, and/or IWA features to predict the *residuals* that would result from applying Ensemble 1 (developed in Step 2) to predict the expert RIASEC ratings for the 269 occupations of interest in this step. Under this strategy, the final model for any given RIASEC dimension would reflect a simple sum of (a) the Ensemble 1 prediction and (b) a residual prediction based on a model consisting of AT, DWA, and/or IWA-based features. Effectively, the Ensemble 1 parameter estimates were treated as fixed in this step, and the only new parameters that were being estimated were those for the residual

prediction models we examined. Treating Ensemble 1 parameter estimates as fixed puts far less demand on the limited sample size available for modeling in this step.

Beyond adopting the residual modeling strategy above, we also changed our sample splitting strategy to make the most use of the 269 occupations available for modeling while still following a prudent cross-validation strategy that properly maintained the independence of training and testing data. Namely, we adopted a nested *k*-fold cross-validation design. We describe this strategy and its implications for the model development process next.

Sample Splitting, Hyperparameter Tuning, and Cross-Validation

Regular *k*-fold cross-validation involves splitting a data set into roughly equal subsamples and systematically setting aside each subsample as a holdout sample on which to evaluate a model trained using the other *k*-1 subsamples. This is the process we applied to our training sample in Step 2 when we tuned hyperparameters; in that design, we also set aside a true holdout sample that we did not analyze until the final stage of model selection. Nested *k*-fold cross-validation adds another layer or "loop" of iteration, such that the design involves both an outer loop and an inner loop of cross-validation analyses. The outer loop of this design is analogous to a regular *k*-fold design, where a holdout sample is defined in each iteration of the analysis and is used to evaluate out-of-sample model performance. The inner loop of this design is where hyperparameter tuning takes place during each iteration of the outer loop: each of the *k*-1 folds of the training sample defined by the outer loop is treated as a holdout sample for evaluating hyperparameter performance. Once the inner loop hyperparameter tuning analyses are complete for a given iteration of the outer loop, the best-performing parameters are used to train a model on the full training subsample (all *k*-1 folds together), and that model is evaluated on the holdout sample defined by the outer loop. The analysis process requires completing a cycle of hyperparameter tuning analyses in the inner loop for each iteration of the outer loop.

To make this design more concrete, Table 6.2 shows how a 5-fold nested cross-validation analysis functions; this is the design we used in the current phase of modeling. In each outer-loop iteration of the analysis, the goal is to use one subsample of data to cross-validate a model trained on the other *k*-1 subsamples. However, since the model to be cross-validated in the outer iteration requires hyperparameter tuning, the inner loop involves subdividing the outer iteration's training sample and performing a separate (*k*-1)-fold cross-validation on that sample to identify the best-performing hyperparameters. Selecting hyperparameters through cross-validation within the inner loop reserves the outer-loop test sample for evaluation of the final model and prevents contaminating the hyperparameter selection process (or the process of comparing machine learning methods applied to the same feature set) with decisions informed by the test data. Thus, the total number of iterations for the outer and inner loops of a nested cross-validation analysis is $k \times (k\text{-}1)$ (i.e., for a 5-fold design, there are 20 total iterations). By using an outer loop of holdout sample analyses rather than establishing a single holdout sample (as we did in our first round of model development research), we made full use of the available data when estimating out-of-sample performance while still maintaining independence of training and test data in each cross-validation analysis.

Table 6.3 shows the sample sizes for all folds of our analysis sample after applying the design shown in Table 6.2, and Table 6.4 shows how job families were represented in each fold of data from that design.

Table 6.2. Five-Fold Nested Cross-Validation Design

Model Cross-Validation (Outer Loop)			Overall Iteration #	Hyperparameter Tuning (Inner Loop)		
Outer Loop Analysis Iteration #	Outer Loop Training Subsample #s	Outer Loop Test Subsample #		Inner Loop Analysis Iteration #	Inner Loop Training Subsample #s	Inner Loop Test Subsample #
1	2, 3, 4, 5	1	1	1	3, 4, 5	2
			2	2	2, 4, 5	3
			3	3	2, 3, 5	4
			4	4	2, 3, 4	5
2	1, 3, 4, 5	2	5	1	3, 4, 5	1
			6	2	1, 4, 5	3
			7	3	1, 3, 5	4
			8	4	1, 3, 4	5
3	1, 2, 4, 5	3	9	1	2, 4, 5	1
			10	2	1, 4, 5	2
			11	3	1, 2, 5	4
			12	4	1, 2, 4	5
4	1, 2, 3, 5	4	13	1	2, 3, 5	1
			14	2	1, 3, 5	2
			15	3	1, 2, 5	3
			16	4	1, 2, 3	5
5	1, 2, 3, 4	5	17	1	2, 3, 4	1
			18	2	1, 3, 4	2
			19	3	1, 2, 4	3
			20	4	1, 2, 3	4

Table 6.3. Sample Breakdown for 5-Fold Nested Cross-Validation

Outer Fold #	Holdout Sample Size (% of Sample)	Total Training Sample size (Size by Inner-Loop Fold)
1	55 (20.4%)	214 (54 / 54 / 53 / 53)
2	54 (20.1%)	215 (55 / 54 / 53 / 53)
3	54 (20.1%)	215 (55 / 54 / 53 / 53)
4	53 (19.7%)	216 (55 / 54 / 54 / 53)
5	53 (19.7%)	216 (55 / 54 / 54 / 53)

Note. Total sample size = 269 O*NET-SOCs.

Table 6.4. Sample Sizes for Job Families Across Data Segments

Job Family		Outer Loop Folds						Inner Loop	
		All	1	2	3	4	5	Training Samples	Folds
11	Management	21	4	4	4	5	4	16 – 17	4 – 5
13	Business and Financial Operations	22	4	5	5	4	4	17 – 18	4 – 5
15	Computer and Mathematical	20	4	4	4	4	4	16 – 16	4 – 4
17	Architecture and Engineering	14	3	3	3	2	3	11 – 12	2 – 3
19	Life, Physical, and Social Science	18	4	4	3	4	3	14 – 15	3 – 4
21	Community and Social Service	2	1	0	0	0	1	1 – 2	0 – 1
23	Legal	3	1	1	1	0	0	2 – 3	0 – 1
25	Educational Instruction and Library	22	5	4	4	4	5	17 – 18	4 – 5
27	Arts, Design, Entertainment, Sports, and Media	19	4	4	3	4	4	15 – 16	3 – 4
29	Healthcare Practitioners and Technical	36	7	7	8	7	7	28 – 29	7 – 8
31	Healthcare Support	3	0	0	1	1	1	2 – 3	0 – 1
33	Protective Service	12	2	3	2	2	3	9 – 10	2 – 3
35	Food Preparation and Serving Related	7	2	1	1	2	1	5 – 6	1 – 2
37	Building and Grounds Cleaning and Maintenance	1	1	0	0	0	0	0 – 1	0 – 1
39	Personal Care and Service	13	2	3	3	3	2	10 – 11	2 – 3
41	Sales and Related	5	1	1	1	1	1	4 – 4	1 – 1
43	Office and Administrative Support	6	1	1	2	1	1	4 – 5	1 – 2
45	Farming, Fishing, and Forestry	4	1	1	0	1	1	3 – 4	0 – 1
47	Construction and Extraction	8	2	1	2	2	1	6 – 7	1 – 2
49	Installation, Maintenance, and Repair	6	1	1	1	1	2	4 – 5	1 – 2
51	Production	11	2	2	3	2	2	8 – 9	2 – 3
53	Transportation and Material Moving	16	3	4	3	3	3	12 – 13	3 – 4
	Total	269	55	54	54	53	53	214 – 216	53 – 55

Note. The number preceding the job family is the first two digits of the O*NET-SOC 2019 code corresponding to that job family.

Residual Model Specifications

We trained our models in two stages that closely resembled the first two modeling stages from Step 2. First, we trained an initial set of three models for each RIASEC dimension according to the specifications in Table 6.5, such that each RIASEC dimension was predicted by separate models based on embeddings from ATs, DWAs, and IWAs, respectively. As noted earlier, in each model, we defined the outcome variable as the residuals obtained from applying Ensemble 1 to the sample of expert ratings collected in Step 5. In other words, the outcome to be predicted for each RIASEC dimension at this stage of our analyses was the difference between the expert ratings and Ensemble 1 predictions.

Table 6.5. Summary of Residual Models to be Trained for Each RIASEC Dimension

Residual Model	Description	Regression Methods Evaluated	# of Features	Feature Type
RM1	Alternate Title SBERT Embeddings	SPLS, EN	768	Alternate Titles
RM2	DWA SBERT Embeddings	SPLS, EN	768	Detailed Work Activities
RM3	IWA SBERT Embeddings	SPLS, EN	768	Intermediate Work Activities

Note. SPLS = Sparse partial least squares regression. EN = Elastic net regression. # of Features = Number of features initially input into the models. SPLS and EN perform variable selection, so the number of features in the final fitted model may be less than the starting number of features initially inputted into the model.

For each training sample in the outer loop of our nested *k*-fold design, we tuned the hyperparameters for elastic net (EN) regression and sparse partial least squares (SPLS) regression models for predicting each RIASEC dimension's residuals from each of the three feature sets. We used a grid search hyperparameter tuning strategy with the same sets of candidate hyperparameter values as we established in Step 2.

Due to our use of a nested *k*-fold design, each combination of RIASEC dimensions and feature sets resulted in five sets of tuned hyperparameters per machine learning method (because each outer loop of our analysis design involved its own hyperparameter tuning process). For this design to produce a final model configuration for each RIASEC dimension and feature set combination, we selected a best-performing method for each analysis based on a vote-counting procedure. For example, if EN regression performed the best in three or more outer folds for a given RIASEC dimension and feature set combination, we considered it the best-bet method for that dimension and feature set; likewise, for SPLS. According to this vote-counting method, EN regression was the best-performing method in 14 out of the 18 RIASEC dimension and feature set combinations. We examined the four instances in which SPLS tended to outperform EN and determined that the advantage gained by using SPLS was truly trivial (the average *RMSE* values across hyperparameter tuning holdout folds were larger for EN than SPLS by magnitudes ranging from .0005 to .0087). Thus, for the sake of parsimony, we chose to use EN as our machine-learning method for all initial models.

Residual Ensemble Specifications

For each EN model we trained on each RIASEC dimension's residuals in each outer loop of our nested *k*-fold design, and we computed predictions for all 269 O*NET-SOCs in our analysis sample. Then, for each RIASEC dimension, we organized the predictions from the AT, DWA,

and IWA models estimated using the same training sample into a data set and trained an ensemble OLS regression model using the AT, DWA, and IWA model predictions as inputs. This analysis approach ensured that all the features used to train in our ensemble models were produced by initial models trained on the same sample of O*NET-SOCs as the ensemble, thus maintaining the independence of all testing and training data segments at all stages of model training.

Cross-Validation and Final Model Fitting

Within each outer loop of the nested *K*-fold design, we trained a residual model on the full training set using the best-functioning set of hyperparameters (if applicable) and evaluated the fit of the model in the corresponding test set. We then averaged the test-set fit metrics across outer-loop analyses for each combination of RIASEC dimensions and feature sets to determine the overall fit of residual models from each instance of the nested *k*-fold analysis.

After identifying the best model/ensemble type to use to predict residuals for each RIASEC dimension, we retrained those models on the full sample of 269 O*NET-SOCs using each outer-loop analysis's best hyperparameter values. For each RIASEC dimension, we then averaged the coefficients for the fully trained residual model/ensemble across outer loop analyses to arrive at a final overall set of coefficients for the RIASEC dimension's best-functioning residual model/ensemble. We then took predictions for those models, added them to the Ensemble 1 predictions, and bounded them by the 1-7 rating metric to arrive at final overall predictions for each RIASEC dimension.

Evaluation of Residual Models and Ensemble

As noted earlier, we experimented with elastic net (EN) regularized regression and sparse partial least squares (SPLS) regression to train our residual models based on aggregated embeddings for ATs, DWAs, and IWAs but, ultimately, we ended up using EN regression for all initial models due to its dominant performance during the hyperparameter tuning process. The hyperparameter values for each EN model are summarized in Table E.1 in Appendix E. We also developed a single OLS regression ensemble model for each RIASEC dimension trained using the predictions from the dimension's AT-, DWA-, and IWA-based residual models.

As in Step 2, we evaluated model fit using root mean squared error (*RMSE*) and multiple *R* metrics but only made decisions based on *RMSE* results because that metric is sensitive to both the strength of linear associations and the correspondence in scale between the target outcome values and model predictions. Tables 6.6 and 6.7 show RMSE and multiple *R* statistics for the three residual models and one residual ensemble we evaluated.

Overall, we found the residual ratings were indeed predictable with a reasonable degree of accuracy, given that the outcome variables in these analyses represented variance that was unaccounted for by Ensemble 1 from Step 2. The average correlation between predicted residuals and observed residuals (i.e., the criterion being modeled) across test set holdout folds hovered in the mid-.4 to .50 range for Realistic and Social interests across models and ranged from high-.3s to high-.4s for Conventional interests. The average holdout correlations were lower for Investigative (high-.2s to high-.3s), Artistic (high-.3s), and Enterprising (mid-.3s to low-.4s) interests.

Table 6.6. Cross-Validated RMSE Results for Residual Models for Each RIASEC Dimension

Model	Average Cross-Validated *RMSE* for Training Data Across Folds								Average Cross-Validated *RMSE* for Test Data Across Folds							
	M		R	I	A	S	E	C	*M*		R	I	A	S	E	C
RM1: Alternate Title SBERT Embeddings	.633		.624	.660	.606	.578	.707	.621	.738		.765	.731	.672	.652	.873	.736
RM2: DWA SBERT Embeddings	.655		.648	.682	.603	.594	.727	.674	.741		.778	.704	.685	.657	.867	.755
RM3: IWA SBERT Embeddings	.671		.678	.676	.604	.588	.784	.693	.746		.783	.709	.675	.655	.882	.772
Ensemble: AT, DWA, and IWA residual predictions	.572		.557	.599	.545	.532	.615	.583	.763		.779	.764	.707	.665	.896	.763

Note. *M* = Average cross-validated *RMSE* across RIASEC dimensions.

Table 6.7. Cross-Validated Multiple R Results for Residual Models for Each RIASEC Dimension

Model	Average Cross-Validated *RMSE* for Training Data Across Folds								Average Cross-Validated *RMSE* for Test Data Across Folds							
	M		R	I	A	S	E	C	*M*		R	I	A	S	E	C
RM1: Alternate Title SBERT Embeddings	.67		.74	.61	.61	.67	.71	.69	.42		.48	.28	.39	.50	.38	.47
RM2: DWA SBERT Embeddings	.61		.70	.50	.58	.61	.65	.60	.42		.47	.39	.37	.48	.38	.44
RM3: IWA SBERT Embeddings	.59		.67	.51	.59	.63	.57	.57	.41		.45	.39	.38	.48	.34	.39
Ensemble: AT, DWA, and IWA residual predictions	.70		.77	.63	.66	.69	.74	.70	.42		.50	.32	.37	.50	.41	.45

Note. *M* = Average cross-validated *R* across RIASEC dimensions.

Upon review of the results in Tables 6.6 and 6.7, we decided to move forward with the residual prediction model based on the alternative title SBERT embeddings (Residual Model 1). This decision was based on both prediction and practical considerations. From a practical perspective, neither DWAs nor IWAs are always immediately available for a new occupation when it is added to O*NET, but alternative titles are. In addition, ATs are the most frequently updated datatype within the O*NET System, potentially serving as an early indicator of change within occupations. From a prediction perspective, the cross-validity of the AT-based residual model was not appreciably different from the others. As such, there was a preference to move forward with the AT-based residual model.

Evaluation of Final RIASEC Predictions: Ensemble 1 + Residual Model 1

Based on the results above, we formed a final prediction for each RIASEC dimension (for each occupation) by summing the Ensemble 1 prediction with the AT-residual model prediction. The 1–7 RIASEC rating scale bounds were then applied to that sum to form a final prediction for each RIASEC dimension (for each occupation). Tables 6.8 and 6.9 show *RMSE* and multiple *R* statistics for final RIASEC predicted values, and Table 6.10 compares the final model performance relative to existing benchmarks.

As shown in Table 6.9, the average test data holdout sample correlations between expert ratings and final additive predictions were in the low-.9s for Realistic, Investigative, Artistic, and Social interests across models. The average holdout sample correlations were slightly lower for Enterprising interests (.87–.88) and considerably lower for Conventional interests (.75–.77). This trend for Conventional interests is consistent with the trends we identified in Step 2, and the results reported by Putka et al. (2023) for their BoW prediction models, and consistent with the lower reliability of Conventional ratings relative to ratings for RIASEC dimensions. Overall, the prediction results were very strong and mirrored the reliability of ratings obtained from trained expert raters (see Table 6.10).

Table 6.8. Cross-Validated RMSE Results for Ensemble 1 + Residual Prediction Models

Model	Average Cross-Validated *RMSE* for Training Data Across Folds								Average Cross-Validated *RMSE* for Test Data Across Folds						
	M		R	I	A	S	E	C	*M*		R	I	A	S	E	C
Ensemble 1	.882		.859	1.049	.722	.771	.903	.987	.896		.845	1.027	.700	.829	.954	1.017
Ensemble 1 + RM1 (Alternate Title)	.618		.613	.666	.568	.554	.668	.640	.718		.743	.731	.629	.622	.833	.747
Ensemble 1 + RM2 (DWA)	.639		.623	.692	.561	.571	.699	.687	.718		.741	.713	.631	.632	.830	.760
Ensemble 1 + RM3 (IWA)*	.656		.652	.690	.567	.567	.752	.705	.726		.749	.720	.628	.632	.847	.778
Ensemble 1 + AT, DWA, and IWA Residual Ensemble	.576		.567	.636	.527	.536	.586	.603	.744		.756	.772	.662	.652	.850	.772

Note. *M* = Average cross-validated *RMSE* across RIASEC dimensions. Results for Ensemble 1 alone are provided as a benchmark for comparison.

Table 6.9. Cross-Validated Multiple R Results for Ensemble 1 + Residual Prediction Models

Model	Average Cross-Validated *R* for Training Data Across Folds								Average Cross-Validated *R* for Test Data Across Folds						
	M		R	I	A	S	E	C	*M*		R	I	A	S	E	C
Ensemble 1	.86		.91	.84	.86	.93	.87	.76	.86		.92	.86	.88	.91	.87	.75
Ensemble 1 + RM1 (Alternate Title)	.91		.95	.93	.92	.94	.92	.83	.89		.92	.92	.91	.93	.88	.77
Ensemble 1 + RM2 (DWA)	.91		.94	.93	.93	.94	.91	.80	.88		.92	.92	.91	.93	.88	.76
Ensemble 1 + RM3 (IWA)*	.90		.94	.93	.92	.94	.90	.79	.88		.92	.92	.91	.93	.87	.75
Ensemble 1 + AT, DWA, and IWA Residual Ensemble	.93		.95	.94	.93	.95	.94	.85	.88		.92	.91	.90	.92	.87	.75

Note. *M* = Average cross-validated *R* across RIASEC dimensions. Results for Ensemble 1 alone are provided as a benchmark for comparison.

Table 6.10. Comparison of Final RIASEC Models' Performance to Existing Benchmarks

Benchmark/Model	*M*		R	I	A	S	E	C
Single Rater Reliability ICC(C,1)	.74		.82	.78	.83	.77	.77	.49
Interrater Reliability ICC(C,3)	.89		.93	.91	.94	.91	.91	.74
Putka et al. (2023) Model Cross-Validated *R*	.84		.90	.83	.83	.92	.85	.73
Final Predictions (Ensemble 1 + Residual Model 1) Average Cross-Validated *R* for Test Data Across Folds	.89		.92	.92	.91	.93	.88	.77

Note. *M* = Average across RIASEC dimensions. ICC(C,1) and ICC(C,3) values are reliability estimates for expert raters in the Step 5 data collection originally reported in Table 5.3.

Additional Evaluations for Final RIASEC Prediction Models

After finalizing our prediction models for each RIASEC dimension and generating predictions based on those models for each of the 269 O*NET-SOCs in the analysis dataset, we conducted follow-up analyses to further evaluate the quality of the predictions. Specifically, we evaluated:

- how well the predicted RIASEC profiles corresponded to the expert rating profiles,
- how well the prediction-implied high-point codes (HPC) corresponded to the expert rating high-point codes,
- patterns of convergence and discrimination among predicted, expert, and analyst RIASEC ratings through a multitrait-multimethod (MTMM) correlation lens,
- the structural validity of the predictions, and
- whether our predictions meaningfully varied in their accuracy by job family or job zone (i.e., whether the distributions of residuals vary by job family or job zone).

These analyses go a step beyond our cross-validated fit estimates and describe our models' performance with greater nuance.

For the first two analyses above, we report the results for three different sample types: training, testing, and full. The training and testing sample results are based on the same models evaluated using the distinct data segments we established earlier. The full sample results are based on the full analysis sample (i.e., 269 occupations) after retraining the residual models using all available cases; these results should be interpreted like the training sample results, as the data used to compute predictions were the same data used to train the models. For the MTMM correlation analyses, we report results based on the stacked set of predictions for the test set holdout samples (i.e., predictions used for each occupation were based on training samples that did not include that occupation) as well as the full sample. For the structural validity analyses, we applied our prediction models to the full set of 923 data-level occupations in O*NET 27.3 and contrasted the results of those analyses to the structural validity of the O*NET 27.3 Database's published RIASEC ratings. Lastly, the residual analyses we report are based solely on the full analysis sample.

RIASEC Profiles

Table 6.11 summarizes the results of our comparisons between O*NET-SOCs' expert-based RIASEC profiles and predicted RIASEC profiles. These results show the predicted profiles show good recovery of expert-based RIASEC profiles, including in the testing samples where predictions were made independent of the data used to train the models.

Table 6.11. Distribution of Within-Occupation RIASEC Profile Correlations and ICCs

Dimension	Training			Testing			Full		
	r	ICC(C,1)	ICC(A,1)	*r*	ICC(C,1)	ICC(A,1)	*r*	ICC(C,1)	ICC(A,1)
N	215.2	215.2	215.2	53.8	53.8	53.8	269	269	269
Mean	.94	.93	.93	.92	.91	.91	.94	.93	.93
SD	.06	.07	.07	.09	.10	.09	.07	.08	.08
Min	.46	.41	.44	.53	.45	.46	.26	.25	.28
5th %ile	.85	.82	.83	.78	.75	.77	.84	.81	.82
10th %ile	.89	.86	.87	.84	.81	.82	.87	.85	.86
20th %ile	.92	.91	.90	.88	.86	.87	.91	.90	.90
30th %ile	.94	.92	.93	.91	.90	.90	.93	.92	.92
40th %ile	.95	.94	.94	.93	.92	.92	.95	.94	.94
Median	.96	.95	.95	.95	.93	.93	.96	.95	.95
60th %ile	.97	.96	.96	.96	.95	.94	.97	.96	.96
70th %ile	.98	.97	.97	.97	.96	.96	.98	.97	.97
80th %ile	.98	.98	.98	.98	.97	.97	.98	.98	.98
90th %ile	.99	.99	.98	.99	.98	.98	.99	.99	.98
95th %ile	.99	.99	.99	.99	.99	.99	.99	.99	.99
Max	1.00	1.00	1.00	1.00	1.00	.99	1.00	1.00	1.00

Note. Results for training and testing samples were averaged over the five outer loops of the nested k-fold design, which is why the sample sizes for those samples include decimals.

High-Point Codes

The results of our HPC analyses based on predicted ratings are summarized in Table 6.12. As in Step 2, we found high rates of agreement between expert-based and predicted HPCs, especially for the first high point and for matches in the first two to three positions when the codes were allowed to appear in any order. A notable difference in these analyses is that we detected overlap in all three HPCs in any order of appearance at higher rates than we detected overlap in the first two HPCs in any order of appearance.

Table 6.12. Agreement on High-Point Codes

High-Point Code Comparison (Predicted vs. Expert)	Training	Testing	Full
1st High-Point	82.8	79.2	82.5
2nd High-Point	52.4	47.5	61.3
3rd High-Point	43.8	39.3	52.4
1st Two High-Points (In Order)	51.0	44.5	60.2
1st Two High-Points (Any Order)	57.8	50.1	68.0
All Three High-Points (In Order)	26.3	24.0	39.4
All Three High-Points (Any Order)	65.6	60.4	69.5
1st High-Point for Predictions Among Top 2 Experts	95.7	93.3	97.0
1st High-Point for Experts Among Top 2 for Predictions	95.6	92.9	96.7
1st High-Point for Predictions Among Top 3 for Experts	97.7	96.6	98.9
1st High-Point for Experts Among Top 3 for Predictions	98.1	96.3	99.3

Note. Results for training and testing samples were averaged over the five outer loops of the nested k-fold design.

Convergence of Predicted Ratings with Analysts and Experts

We examined predicted, analyst, and expert RIASEC ratings through a multitrait-multimethod (MTMM) correlation lens to evaluate patterns of convergence and discrimination among ratings. The purpose of doing so was to establish convergent and discriminant validity (Campbell & Fiske, 1959). Tables 6.13 and 6.14 provide MTMM correlations based on (a) the stacked set of predictions for the test set holdout samples (Table 6.13), as well as (b) predictions based on the full sample. The results in Table 6.13 avoid capitalizing on chance, as the predictions for each occupation used to calculate the MTMM correlations were based on training samples that did not include the given occupation.

Convergent validity for predicted RIASEC ratings is indicated by high correlation among the same RIASEC dimension based on predictions, analysts, and experts (i.e., monotrait-heteromethod correlations, highlighted in yellow in Tables 6.13 and 6.14). Focusing on results in Table 6.13 to avoid capitalizing on chance, we see strong evidence of convergence between predictions and analyst ratings and predictions and expert ratings. The average monotrait-heteromethod correlation among predicted and analyst ratings was. 82, whereas among predicted and expert ratings it was .88. Predicted-expert and predicted-analyst monotrait-heteromethod correlations for RIASEC dimensions all exceed .80 except for Conventional interests (predicted-expert r = .76, predicted-analyst r = .67). Note that when interpreting the predicted-expert and predicted-analyst correlation for Conventional interests, it is important to remember that this dimension was the least reliably measured among both experts and analysts (recall Table 5.3). When we correct these correlations for unreliability in expert and analyst ratings using the ICC(C,3) (expert) and ICC(C,6) (analyst) values for Conventional interests in Table 5.3, they increase to corrected values of .75 (predicted-analyst), and .88 (predicted-expert).

.

Table 6.13. Multitrait-Multimethod Correlations for RIASEC Dimensions by Rating Source: Stacked Predictions for Test Set Holdouts

		Analyst						Expert						Predicted					
		R	I	A	S	E	C	R	I	A	S	E	C	R	I	A	S	E	C
Analyst	R	1.00																	
	I	-.26	1.00																
	A	-.04	-.39	1.00															
	S	-.49	-.02	.16	1.00														
	E	-.35	-.26	.00	-.02	1.00													
	C	-.49	.31	-.33	-.12	.31	1.00												
Expert	R	.84	-.06	-.09	-.35	-.44	-.47	1.00											
	I	-.37	.81	-.19	.04	-.10	.44	-.17	1.00										
	A	-.18	-.25	.85	.14	.12	-.20	-.24	-.03	1.00									
	S	-.43	.08	.14	.85	-.09	-.06	-.36	.16	.13	1.00								
	E	-.35	-.24	.02	.18	.85	.19	-.45	-.14	.10	.07	1.00							
	C	-.26	.20	-.34	-.29	.12	.61	-.25	.14	-.32	-.33	.00	1.00						
Predicted	R	.84	-.07	-.11	-.38	-.46	-.48	.92	-.19	-.24	-.38	-.47	-.21	1.00					
	I	-.36	.83	-.21	-.01	-.15	.44	-.16	.91	-.07	.14	-.18	.17	-.17	1.00				
	A	-.19	-.27	.85	.16	.10	-.20	-.25	-.01	.90	.17	.09	-.33	-.28	-.03	1.00			
	S	-.48	.11	.10	.86	-.09	-.08	-.40	.15	.10	.92	.09	-.30	-.43	.14	.15	1.00		
	E	-.36	-.27	.02	.17	.85	.24	-.46	-.18	.09	.07	.87	.05	-.50	-.25	.09	.09	1.00	
	C	-.31	.12	-.40	-.30	.24	.67	-.31	.06	-.36	-.37	.09	.76	-.31	.08	-.41	-.36	.14	1.00

Note. N = 269. Monotrait-heteromethod correlations are highlighted in yellow. Heterotrait-monomethod correlations are highlighted in blue. Heterotrait-heteromethod correlations are highlighted in green.

Table 6.14. Multitrait-Multimethod Correlations for RIASEC Dimensions by Rating Source: Predictions for Full Sample

		Analyst						Expert						Predicted					
		R	I	A	S	E	C	R	I	A	S	E	C	R	I	A	S	E	C
Analyst	R	1.00																	
	I	-.26	1.00																
	A	-.04	-.39	1.00															
	S	-.49	-.02	.16	1.00														
	E	-.35	-.26	.00	-.02	1.00													
	C	-.49	.31	-.33	-.12	.31	1.00												
Expert	R	.84	-.06	-.09	-.35	-.44	-.47	1.00											
	I	-.37	.81	-.19	.04	-.10	.44	-.17	1.00										
	A	-.18	-.25	.85	.14	.12	-.20	-.24	-.03	1.00									
	S	-.43	.08	.14	.85	-.09	-.06	-.36	.16	.13	1.00								
	E	-.35	-.24	.02	.18	.85	.19	-.45	-.14	.10	.07	1.00							
	C	-.26	.20	-.34	-.29	.12	.61	-.25	.14	-.32	-.33	.00	1.00						
Predicted	R	.85	-.06	-.11	-.37	-.47	-.48	.94	-.18	-.25	-.37	-.48	-.22	1.00					
	I	-.36	.83	-.21	.00	-.15	.44	-.16	.93	-.06	.14	-.18	.17	-.17	1.00				
	A	-.20	-.27	.86	.17	.11	-.20	-.25	-.01	.92	.18	.09	-.34	-.27	-.03	1.00			
	S	-.48	.11	.11	.86	-.09	-.08	-.40	.16	.11	.94	.09	-.30	-.43	.14	.15	1.00		
	E	-.36	-.28	.02	.17	.86	.23	-.47	-.18	.09	.06	.92	.04	-.50	-.24	.09	.08	1.00	
	C	-.31	.13	-.40	-.32	.22	.69	-.31	.07	-.37	-.37	.07	.82	-.31	.09	-.42	-.36	.13	1.00

Note. N = 269. Monotrait-heteromethod correlations are highlighted in yellow. Heterotrait-monomethod correlations are highlighted in blue. Heterotrait-heteromethod correlations are highlighted in green.

Discriminant validity and freedom from "common method" variance are indicated by relatively low correlations among different RIASEC dimensions that had a rating source in common (i.e., heterotrait-monomethod correlations, highlighted in blue in Tables 6.13 and 6.14) and that do not clearly exceed the correlations among different RIASEC dimensions evaluated by different sources (i.e., heterotrait-heteromethod correlations, highlighted in green in Tables 6.13 and 6.15). Again, focusing on the results in Table 6.13, we see a pattern of evidence for discriminant validity and little evidence of common method variance. Specifically, the average heterotrait-monomethod correlation among predicted ratings was: -.14. Among analyst ratings, it was -.13, and among expert ratings, it was -.11. In comparison, the average heterotrait-heteromethod correlation (across all sources) was -.11.

Structural Validity

Several analyses were conducted to examine the structural validity of the new predicted RIASEC ratings. For comparison purposes, we also examined the structural validity of the RIASEC ratings published in the O*NET 27.3 database. The structural validity tests are based on Holland's (1997) hexagonal model (also called a circular-order model). A randomization test of hypothesized order (Rounds et al., 1992; Tracey, 1997) was conducted on the correlation matrix of the predicted and published ratings. A spatial analysis using multidimensional scaling (MDS) was conducted to display the inter-relations among the RIASEC ratings. The MDS analyses were conducted using the *smacof* R package (Version 2.1-5; Mair et al., 2022).

Table 6.15 displays the RIASEC intercorrelations for new predicted ratings (n = 923 occupations) and published O*NET 27.3 ratings (n = 874 occupations). Because of the circular order of Holland's RIASEC model, it is expected that the correlations decrease as one scale moves farther away from other RIASEC scales around a circular structure. For example, Realistic should be more highly correlated with Investigative compared to Artistic. Overall, the circular order correlation pattern holds for the new predicted ratings and published O*NET 27.3 ratings.

Table 6.15. RIASEC Intercorrelations based on New Predicted RIASEC Ratings and Published O*NET 27.3 RIASEC Ratings

RIASEC Correlations based on New Predicted Ratings (n = 923 occupations)								RIASEC Correlations based on Published O*NET 27.3 Ratings (n = 874 occupations)						
	R	I	A	S	E	C			R	I	A	S	E	C
R	1.00	-.15	-.33	-.59	-.69	-.32		R	1.00	-.07	-.39	-.58	-.56	-.13
I	-.15	1.00	.12	.23	-.16	-.07		I	-.07	1.00	.20	.07	-.30	-.17
A	-.33	.12	1.00	.32	.09	-.47		A	-.39	.20	1.00	.32	.02	-.41
S	-.59	.23	.32	1.00	.27	-.26		S	-.58	.07	.32	1.00	.19	-.24
E	-.69	-.16	.09	.27	1.00	.30		E	-.56	-.30	.02	.19	1.00	.27
C	-.32	-.07	-.47	-.26	.30	1.00		C	-.13	-.17	-.41	-.24	.27	1.00

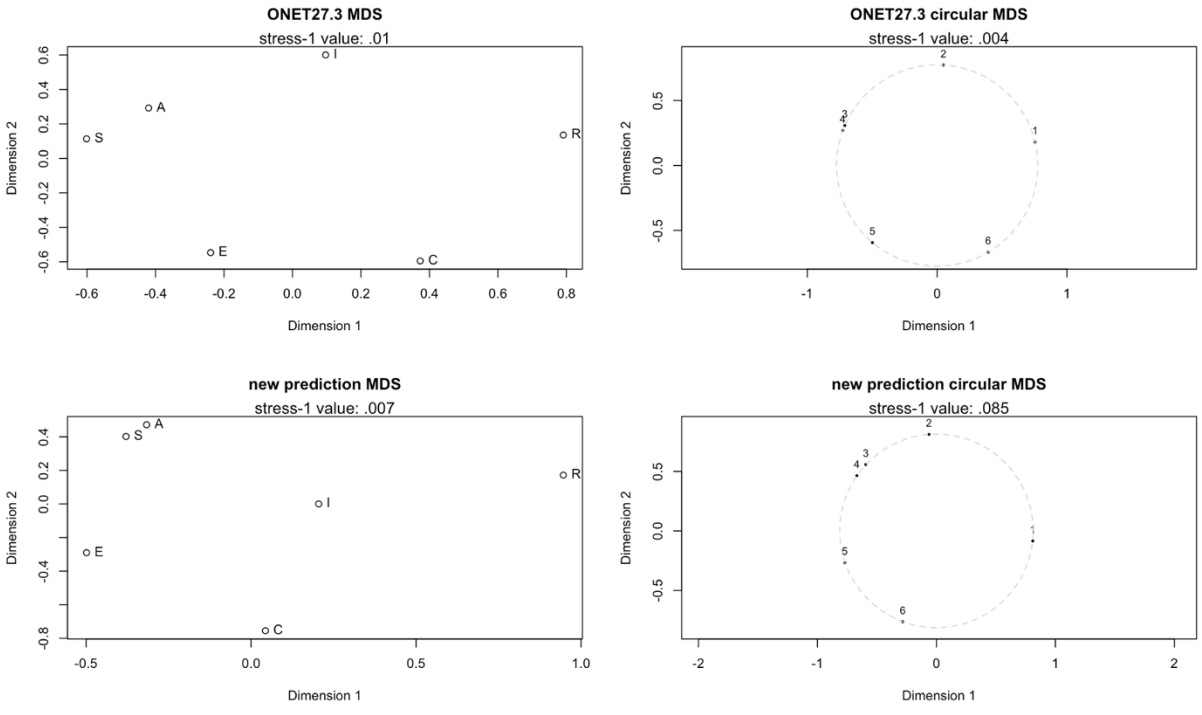
Note. Correlations are shaded along a green-red color gradient to facilitate interpretation (higher values are shaded green, lower values are shaded red).

Results from the randomization test are twofold. First, the test evaluates RIASEC-ordered correlations to determine whether they are random given the RIASEC circular order structure. Second, the correspondence index (*CI*) is reported. The *CI* is a normalized descriptive statistic indicating the degree to which the 72 ordered predictions implied by the RIASEC circular order structure (e.g., Realistic is more highly correlated with Investigative than Artistic) are met. The

CI varies from -1 to 1, with positive values indicating more of the 72 predictions are met and 0 indicating chance agreement or disagreement. For both data sets, the random order hypothesis was rejected. The *CI* for the predicted ratings was .61 ($p < .01$), and for the published O*NET 27.3 ratings it was .78 ($p < .01$). Both *CI* values indicate that the RIASEC correlations fit Holland's model as well as the U.S. benchmark sample (*CI* = .70; Rounds & Tracey, 1996), the Interest Profiler Short Form *CI* of .69, and the Paper and Pencil form *CI* of .61 (Rounds et al., 2021).

Figure 6.1 plots MDS and constrained solutions for the new predicted ratings and published O*NET 27.3 ratings. A two-dimensional solution fits the data well. MDS solutions for both the new predicted and published ratings (plots on the left-hand side of Figure 6.1) show the distance between Realistic interests and Investigative and Conventional interests is greater than would be expected given a circular structure, which is a typical finding in the RIASEC structural literature (Rounds & Day, 1999). Comparing the predicted and published rating MDS solutions shows that the major difference between them is that Investigative interests for the predicted ratings are drawn closer to the center of the configuration. The circular structure of the RIASEC ratings is further supported by the constrained (circular MDS) scaling results (plots on the right side of Figure 6.1). Notably, both datasets allowed for the placement of interest types on a circular configuration with good stress values for the new predicted ratings (stress = .085) and published O*NET 27.3 ratings (stress = .004).

*Figure 6.1. Multidimensional Scaling and Constrained (Circular MDS) Solution Plots for Predicted Ratings and Published O*NET 27.3 Ratings*



Note. R,1 = Realistic, I,2 = Investigative, A,3 = Artistic, S,4 = Social, E,5 = Enterprising, C,6 = Conventional.

Residual Analyses

We computed residuals (final predicted ratings minus expert ratings) for each RIASEC dimension for each of the 269 occupations in our analysis sample, and we examined the distributions of these residuals across O*NET job families and job zones. The purpose of these analyses was to determine whether the quality of predictions afforded by the RIASEC prediction models varied meaningfully by job family or job zone.

To evaluate the extent to which variation in residuals was attributable to job family or job zone-level differences as opposed to occupation-level differences, we decomposed the variance in residuals using two pairs of simple random effects models. The first pair of models treated occupations as nested within job families and decomposed variance in raw residuals and absolute differences between expert and predicted ratings (i.e., the absolute value of the raw residual) into job family and occupation components. The second pair of models treated occupations as nested within job zones and decomposed variance in raw residuals and absolute differences between expert and predicted ratings into job zone and occupation components. The results of these analyses are presented in Tables 6.16 and 6.17.

Table 6.16. Percentage of Variance in Prediction Residuals Attributable to Job Family vs. Occupation

Dimension	Raw Residual			Absolute Residual	
	% Job Family	% Occupation		% Job Family	% Occupation
Realistic	0.0	100.0		3.7	96.3
Investigative	4.0	96.0		3.9	96.1
Artistic	10.3	89.7		19.8	80.2
Social	4.4	95.6		6.0	94.0
Enterprising	0.0	100.0		0.3	99.7
Conventional	16.3	83.7		7.1	92.9

Note. Cell values reflect percentages of variance in residual and absolute residuals across O*NET-SOCs (n = 269) attributable to the given factor (job family or occupation) based on restricted maximum likelihood (REML) variance components.

Table 6.17. Percentage of Variance in Prediction Residuals Attributable to Job Zone vs. Occupation

Dimension	Raw Residual			Absolute Residual	
	% Job Zone	% Occupation		% Job Zone	% Occupation
Realistic	0.0	100.0		0.8	99.2
Investigative	0.0	100.0		1.5	98.5
Artistic	1.2	98.8		2.5	97.5
Social	0.0	100.0		4.3	95.7
Enterprising	2.8	97.2		0.0	100.0
Conventional	2.5	97.5		1.2	98.8

Note. Cell values reflect percentages of variance in residual and absolute residuals across O*NET-SOCs (n = 269) attributable to the given factor (job zone or occupation) based on restricted maximum likelihood (REML) variance components.

Results presented in Tables 6.16 and 6.17 suggest that the quality of prediction for Realistic, Investigative, Social, and Enterprising interests did not appear to vary much across job families or job zones, as these factors accounted for a relatively low percentage (less than 7%) of variance in raw residuals and absolute residuals across O*NET-SOCs. For Artistic and Conventional interests, the quality of predictions did not appear to vary much across job zones, as it accounted for no more than 2.5% of the variance in raw residuals and absolute residuals across O*NET-SOCs. However, the quality of predictions did appear to vary more notably for Artistic and Conventional interests across job families. Specifically, job family accounted for 10.3% of the variance in raw residuals for Artistic (19.8% of the variance in absolute residuals) and 16.3% of the variance in raw residuals for Conventional (7.1% of the variance in absolute residuals).

To better understand the nature of these differences Appendix F provides means and standard deviations of raw and absolute residuals by job family (Table F.1) and job zone (Table F.2). Focusing on the job family results in Table F.1 and F.2 (as that is where the more meaningful differences occurred for Artistic and Conventional interests) and those job families that are represented by at least five occupations, we observed that our final prediction model for Artistic interests tended to underpredict how descriptive Artistic interests were for Arts, Design, Entertainment, Sports and Media occupations (n = 19) by an average of .62 rating scale units relative to experts, and tended to overpredict how descriptive Artistic interests were for Production occupations (n = 11) by an average of .50 rating scale units relative to experts (see Table F.1).[16] Furthermore, we observed that our final prediction model for Conventional interests tended to overpredict how descriptive Conventional interests were for Food Preparation and Serving Related occupations (n = 7) and Sales and Related occupations (n = 5) by an average of .64 and .74 rating scale units, respectively. We caution the reader against overinterpreting the findings summarized above, in that the residual summaries for job families presented in Tables F1 and F.2 are based on relatively small numbers of occupations per job family. Given the strength of the findings presented for our final predictions in the previous sections, the totality of evidence suggests our prediction models perform well in terms of their alignment with both expert and analyst ratings. We, however, suggest the Center use the residuals summaries in Appendix F to perhaps give closer scrutiny to occupations from families with larger residuals for a given RIASEC dimension when these models are used to update ratings in future versions of O*NET.

[16] Note the choice of five occupations here as a minimum is somewhat arbitrary and chosen simply to facilitate discussion of pattern of findings in Tables F.1 and F.2.

Step 7: Finalizing OIPs and High-Point Codes for O*NET 28.1

Upon conclusion of Step 6, we applied our final prediction model for each RIASEC dimension (i.e., Ensemble 1 + Residual Model 1) to the 923 data-level occupations in O*NET 27.3 (i.e., the O*NET database that was most current as of this step in the research effort) to generate RIASEC predictions for each occupation. Based on ratings, we then assigned up to three high-point codes for each occupation using the three steps outlined earlier under Step 3.

With these ratings and high-point codes in place, we constructed a file consisting of (a) the aforementioned predicted ratings and high-point codes for all 923 occupations, (b) final expert ratings and high-point codes for the subset of 269 occupations that had them, and (c) published O*NET 27.3 database RIASEC ratings and 874 occupations that had high-point codes. The purpose of constructing this file was to facilitate a final review of predicted RIASEC ratings and high-point codes for the occupations. Dr Round adopted the following three steps for purposes of review:

- **Step 1.** Identified occupations that had RIASEC profile correlations less than .80 for predicted and expert ratings or between predicted and published ratings (as of O*NET 27.3). The rationale behind this step was that Holland's theory is grounded in the concept of person-environment fit. As such, profile correlations serve as measures of fit, congruence, and correspondence. Profile correlations are used to identify occupations (OIPs) that fit individual RIASEC interests (Gregory & Lewis, 2016). Differences between profile correlations generated by predicted and expert ratings can result in different occupations that match an individual's interests.

- **Step 2**. For occupations identified in Step 1 (i.e., profile correlations less than .80), high-point RIASEC codes were examined more closely. First, Dr. Rounds identified the occupations that did not have matching RIASEC dimensions in the first position of the high-point codes based on predicted and expert ratings (and predicted and published ratings). For occupations that did not meet this first position match criterion, Dr. Rounds then examined whether the RIASEC dimension in the first or second position of the high-point codes matched based on predicted and expert ratings (and predicted and published ratings). The rationale behind this step is that many methods of fitting individual RIASEC interests to occupations have used one or two high-point codes. Clients and practitioners can use the RIASEC high-point codes to identify occupations that fit well.

- **Step 3**. The three-point RIASEC codes for occupations flagged via the steps above were then carefully examined by Dr. Rounds.

Results of Review

Predicted vs. Expert Ratings

Of the 269 occupations that had both predicted and expert ratings, Dr. Rounds identified 13 occupations that fell below the .80 RIASEC profile correlation threshold, none of which had matching RIASEC dimensions in the first high-point code position based on predicted and expert ratings. Nevertheless, all these occupations did have matching first or second position high-point codes based on predicted and expert ratings (after breaking ties on the expert side). Further review of these occupations supported the use of the predicted ratings and high-point codes for them.

Predicted vs. Published Ratings

Of the 874 occupations that had both predicted and published ratings, Dr. Rounds identified 66 occupations that fell below the .80 RIASEC profile correlation threshold. Of these 66 occupations, 37 had mismatching RIASEC dimensions in the first high-point code position based on predicted and published ratings. Of the aforementioned 37 occupations, only the following six did not have matching first or second position high-point codes based on predicted and published ratings:

- 15-1231.00: Computer Network Support
- 29-2032.00: Diagnostic Medical
- 33-9091.00: Crossing Guards and Flaggers
- 39-5012.00: Hair, Hairstylists, and Cosmetologists
- 39-3093.00: Locker Room, Coatroom, and Dressing Room Attendants
- 53-2031.00: Flight Attendants

Upon further review of these occupations, it was decided to adopt the predicted ratings and high-point codes for the occupations as they appeared consistent with the tasks performed in these occupations.

Summary

Based on the results of Dr. Rounds's review, no changes were deemed necessary to the predicted ratings or predicted high-point codes. For occupations where differences were found, Dr. Rounds leaned towards using the predicted ratings based on further review of the occupations. The findings here reinforce the notion that actuarial (statistical) judgments can offer more accurate judgments than those clinical judgments made by humans (Dawes et al., 1989; Grove & Meehl, 1996).

In sum, based on the process above, we arrived at a final set of RIASEC ratings and high-point codes for the 923 data-level O*NET-SOCs based on inputs available in O*NET 27.3. The final set of RIASEC ratings and high-point codes produced here is expected to first be published in O*NET 28.1 targeted for release in November 2023.

Guidance for Updating RIASEC Ratings and High-Point Codes in Future Versions of the O*NET Database

To facilitate the application of our final prediction models to future versions of O*NET database as new occupations are added or as existing occupation data changes, we combined our final feature generation and prediction model code into a single script that takes an O*NET Database folder as its input and produces interest predictions and high-point codes as its output in a format that mimics the structure of the Interests table in the O*NET Database. The aforementioned O*NET Database folder would include all tables from the future version of the O*NET Database that contribute inputs to our final prediction models, namely:

- Occupation Data
- Task Statements
- Alternate Titles

Our suggested updating process is as follows:

1. Download the aforementioned tables from the latest available version of the O*NET database.
2. Run the aforementioned feature-generation and model prediction code to produce a draft updated version of the Interests table.
3. Run code that creates occupation-level binary variables indicating which, if any, of the input files (from the bulleted list above) have changed for each occupation between the version of the O*NET database *upon which the latest available interest data were created* and the latest available version of the O*NET database. For example, if the next time interest data are updated is for O*NET 29.0, the version of the O*NET database *upon which the latest available interest data were created* would be 27.3 (per Step 7 in this report), and the latest available version of the O*NET database would be 29.0. These binary variables described here will be added to the file merged file created as part of the next step.
4. Run code that merges the updated version of the Interest table (from #2) and the Interest table from the latest available version of the O*NET Database and creates the following occupation-level flags to identify those occupations where closer review of updated data by a RIASEC expert is most critical:

    a. Occupations where the correlation between RIASEC profile based on the updated ratings vs. the latest available ratings is < .80.

    b. Occupations where the first high-point code does not match based on the latest available ratings.

    c. Occupations that have no interest ratings or high-point codes in the Interest table from the latest available version of the O*NET Database (i.e., new occupations).

    The flags above are meant to focus an expert's review on those occupations where the set of updated RIASEC data for an occupation is least aligned with the latest available RIASEC data for that occupation. Within the set of occupations flagged above, the

expert could limit their review to those occupations where inputs changed between the versions of the O*NET database on which the updated and latest available interest data are based. The RIASEC expert tasked with providing a final review of updated ratings before they go "live" to the operational database would make adjustments to the ratings/high-point codes as needed based upon their review (similar to the final review conducted as part of Step 7 in the current effort).

5. Following RIASEC expert review and revision, the final updated Interest table is provided to the Center for final review prior to being incorporated into the operational O*NET Database.

Timing of Future Interest Rating and High-Point Code Updates

Ultimately, the frequency with which the Center adopts the updating process above will come down to its priorities and timeliness with which it wishes to address changes that may impact the currency of interest data in O*NET. As an example, consider the following two extremes in terms of updating frequency:

- The Center may consider making updates to interest data (OIPs and high-point codes) *every time* information for an occupation that serves as input to the RIASEC prediction models changes (i.e., occupation title, description, tasks, alternate titles), or a new occupation is introduced into the O*NET Database. The update process above could be carried out following the introduction of those new inputs or new occupations, and interest data for the occupations in question would be incorporated into a subsequent version of the O*NET Database.
- The Center may consider making updates to interest data database-wide every *X* years (e.g., every two years) for all occupations that had input that changed in that time frame or for any new occupations that were introduced in that time frame. The update process above could be carried out at a fixed point in time every *X* years, and interest data for the occupations in question would be incorporated into a subsequent version of the O*NET Database following that time frame.

The ideal updating solution may depend on how often new occupations are introduced to O*NET and how often the inputs into the RIASEC prediction model (i.e., occupation title, description, tasks, alternate titles) change for existing occupations. In discussing potential updating options for existing occupations with the Center, it was noted among the inputs above, most changes would likely be to alternate titles. The occupation title, description, and task statement lists historically have been fairly stable, with the most notable changes stemming from (a) adjusted core/supplemental task designations, (b) task statements removed from task listings, and (c) modified occupation descriptions occurring during an occupational taxonomy update. In the future, a new emphasis on identifying emerging tasks may lead to an increased frequency in which new task statements are added to listings in between the Interest ratings updates.

Conclusions and Future Directions

The present report summarized the successful effort to leverage advances in supervised machine learning to populate RIASEC OIPs and high-point codes for 923 data-level O*NET-SOC occupations. The models developed use readily available information published within the O*NET database as input for generating OIPs and high-point codes for current and new occupations. As the world of work changes, these models can be applied to future versions of the O*NET database to generate and maintain high quality vocational interest information for the O*NET System.

For each of the six RIASEC general interests, we developed models that produced predictions that converged well with both expert and O*NET analyst ratings and the level of correlations we observed between predicted and expert approached levels of interrater reliability seen among expert raters. The developed models will eliminate the need to gather OIPs and high-point codes via traditional expert or analyst data collections, helping to ensure updated, accurate vocational interests information is available on a timely basis for O*NET customers and stakeholders.

Though our prediction models performed well, we still advise the Center to build in a layer of expert review of predicted OIPs and high-point codes when they are generated for new O*NET occupations or for occupations where key model inputs have changed. As part of this report, we've provided a recommended updating and expert review process.

In terms of future directions, the success of the models evaluated here bodes well for potentially extending this work to automate the creation of basic interest profiles for all data-level occupations. The Center recently undertook an initiative to add basic interests to the O*NET Content Model (see Rounds et al., 2023) but is still in need of a method for efficiently profiling occupations on those basic interests. The methods examined here suggest a machine learning approach could be a promising avenue to pursue for this purpose.

References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105. https://doi.org/10.1037/h0046016

Chun, H., & Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 72*(1), 3–25. https://doi.org/10.1111/j.1467-9868.2009.00723.x

Dahlke, J. A., & Putka, D. J. (2022). Automating the generation of vocational interest profiles from occupational descriptions. In B. P. Acton, & N. Koenig (Chair), *Automating is the future: Improving research scalability with predictive modeling* (Symposium). 2022 Annual Society for Industrial and Organizational Psychology Conference, Seattle, WA.

Dahlke, J. A., Putka, D. J., Shewach, O., & Lewis, P. (2022). *Developing related occupations for the O*NET Program.* National Center for O*NET Development. https://www.onetcenter.org/reports/Related_2022.html

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. ArXiv:1810.04805 [Cs]. http://arxiv.org/abs/1810.04805

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33(1)*, 1–22.

Gregory, C., & Lewis, P. (2016). *Linking client assessment profiles to O* NET® occupational profiles within the O* NET Interest Profiler Short Form and Mini Interest Profiler (Mini-IP)*. National Center for O*NET Development. https://www.onetcenter.org/reports/Mini-IP_Linking.html

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

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2(2)*, 293–323. https://doi.org/10.1037/1076-8971.2.2.293

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R, 2nd edition*. Springer.

Mair, P., Groenen, P. J. F., & De Leeuw, J. (2022). More on multidimensional scaling in R: smacof version 2. *Journal of Statistical Software*, *102*(10), 1-47. https://doi.org/10.18637/jss.v102.i10

National Center for O*NET Development (2023). *O*NET 28.0 Database*. O*NET Resource Center. https://www.onetcenter.org/database.html

Putka, D. J., Oswald, F. L., Landers. R. N., McCloy, R. A., Beatty, A. S., & Yu, M. C. (2023). Evaluating a natural language processing approach to estimating KSA and interest job analysis ratings. *Journal of Business and Psychology, 38*, 385-410. https://link.springer.com/article/10.1007/s10869-022-09824-0

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. arXiv:1908.10084. 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*. National Center for O*NET Development. https://www.onetcenter.org/reports/JobZoneProcedureUpdate.html

Rounds, J., Armstrong, P. I., Liao, H. Y., Lewis, P., & Rivkin, D. (2008). *Second generation occupational interest profiles for the O* NET system: Summary*. National Center for O*NET Development. https://www.onetcenter.org/reports/SecondOIP_Summary.html

Rounds, J., & Day, S. X. (1999). Describing, evaluating, and creating vocational interest structures. In M. L. Savickas & A. R. Spokane (Eds.), *Vocational interests: Meaning, measurement, and counseling use* (pp. 103–133). Davies-Black Publishing.

Rounds, J., Hoff, K., & Lewis, P. (2021). *O*NET Interest Profiler Manual*. National Center for O*NET Development. https://www.onetcenter.org/reports/IP_Manual.html

Rounds, J., Putka, D. J., & Lewis, P. (2023). *Updating vocational interest information for the O*NET Content Model*. National Center for O*NET Development. https://www.onetcenter.org/reports/Voc_Interests.html

Rounds, J., Smith, T., Hubert, L., Lewis, P., & Rivkin, D. (1999). *Development of Occupational Interest Profiles (OIPs) for the O*NET*. National Center for O*NET Development. https://www.onetcenter.org/reports/OIP.html

Rounds, J., Su, R., Lewis, P., & Rivkin, D. (2010). O*NET Interest Profiler Short form psychometric characteristics: Summary. National Center for O*NET Development. http://www.onetcenter.org/reports/IPSF_Psychometric.html

Rounds, J., Su, R., Lewis, P., & Rivkin, D. (2013). *Occupational interest profiles for new and emerging occupations in the O* NET system: Summary*. National Center for O*NET Development. https://www.onetcenter.org/reports/OIP_NewEmerging.html

Rounds, J., & Tracey, T. J. (1996). Cross-cultural structural equivalence of RIASEC models and measures. *Journal of Counseling Psychology, 43*(3), 310–329. https://doi.org/10.1037/0022-0167.43.3.310

Rounds, J., Tracey, T. J., & Hubert, L. (1992). Methods for evaluating vocational interest structural hypotheses. *Journal of Vocational Behavior*, *40*(2), 239–259. https://doi.org/10.1016/0001-8791(92)90073-9

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Su, R., Tay, L., Liao, H-Y, Zhang, Q., & Rounds, J. (2019). Toward a dimensional model of vocational interests. *Journal of Applied Psychology*, *104*, 690-714. https://doi.org/10.1037/apl0000373

Tracey, T. J. G. (1997). RANDALL: A Microsoft FORTRAN program for a randomization test of hypothesized order relations. *Educational and Psychological Measurement, 57*(1), 164–168. https://doi.org/10.1177/0013164497057001012

Appendix A: RIASEC Dimension Descriptions from the O*NET Content Model

The RIASEC dimension descriptions presented below are from the O*NET Content Model Reference table updated in O*NET 28.0.

Table A.1. RIASEC Dimension Descriptions from the O*NET Content Model

Dimension	Description
Realistic	Work involves designing, building, or repairing of equipment, materials, or structures, engaging in physical activity, or working outdoors. Realistic occupations are often associated with engineering, mechanics and electronics, construction, woodworking, transportation, machine operation, agriculture, animal services, physical or manual labor, athletics, or protective services.
Investigative	Work involves studying and researching non-living objects, living organisms, disease or other forms of impairment, or human behavior. Investigative occupations are often associated with physical, life, medical, or social sciences, and can be found in the fields of humanities, mathematics/statistics, information technology, or health care service.
Artistic	Work involves creating original visual artwork, performances, written works, food, or music for a variety of media, or applying artistic principles to the design of various objects and materials. Artistic occupations are often associated with visual arts, applied arts and design, performing arts, music, creative writing, media, or culinary art.
Social	Work involves helping, teaching, advising, assisting, or providing service to others. Social occupations are often associated with social, health care, personal service, teaching/education, or religious activities.
Enterprising	Work involves managing, negotiating, marketing, or selling, typically in a business setting, or leading or advising people in political and legal situations. Enterprising occupations are often associated with business initiatives, sales, marketing/advertising, finance, management/administration, professional advising, public speaking, politics, or law.
Conventional	Work involves following procedures and regulations to organize information or data, typically in a business setting. Conventional occupations are often associated with office work, accounting, mathematics/statistics, information technology, finance, or human resources.

Appendix B: 2008-2013 Interest Re-Rating Instructions

Instructions: This workbook contains four tabs labeled “Rating Instructions”, “Rating Sheet”, “Descriptions” and “Task Statements.” The Rating Sheet tab is where you’ll make ratings for those occupations-by-RIASEC combination where we need ratings and are sorted by O*NET-SOC first (ascending) and RIASEC dimension to be re-rated second (ascending). The Descriptions tab included descriptions for each occupation where ratings are needed (sorted in ascending order of O*NET-SOC). The Task Statements tab includes tasks along with their associated importance ratings (sorted in ascending order of O*NET-SOC, then in descending order of importance).

This activity involves reviewing and resolving ratings for occupation-by-RIASEC combinations flagged for disagreement – note some occupations will require re-rating on multiple RIASEC dimensions. Suggested steps for re-rating:

1. Print out the Description and Task Statement tabs. Both have been formatted for easy printing.

2. With the Descriptions and Task Statements in hand, open the Rating Sheet and provide a new “characteristicness/descriptiveness” rating for the RIASEC dimension listed in the column named “RIASEC Dimension to be Re-Rated”. Enter your rating in the column named “Dr. Rounds rating”.

3. To facilitate making your ratings, we’ve provided ratings from the original three SMEs who rated the occupation-dimension in question.

4. When making your ratings please use the following scale:

“Rate the Occupational Units on each of the RIASEC work environments using the following seven point scale. Ask yourself, ‘How descriptive and characteristic is the Holland work environment of this Occupational Unit?’”

Not at all characteristic			Moderately characteristic			Extremely characteristic
1	2	3	4	5	6	7

Appendix C: Best Performing Regression Method and Hyperparameter Values by Model and Ensemble

Table C.1. Best-Performing Machine Learning Methods and Hyperparameter Values for Initial Models

Model	Realistic	Investigative	Artistic	Social	Enterprising	Conventional
1	SPLS (K = 3, η = 0.1)	SPLS (K = 2, η = 0.0)	SPLS (K = 3, η = 0.0)	SPLS (K = 3, η = 0.0)	SPLS (K = 2, η = 0.0)	SPLS (K = 2, η = 0.2)
2	EN (α = 0.0, λ = 1.9630)	EN (α = 0.0, λ = 3.9442)	EN (α = 0.0, λ = 1.2328)	EN (α = 0.0, λ = 3.5112)	EN (α = 0.0, λ = 1.9630)	EN (α = 0.0, λ = 3.9442)
3	EN (α = 0.0, λ = 3.9442)	EN (α = 0.0, λ = 3.9442)	EN (α = 0.0, λ = 2.2051)	EN (α = 0.0, λ = 3.9442)	EN (α = 0.0, λ = 1.2328)	EN (α = 0.0, λ = 4.4306)
4	EN (α = 0.0, λ = 2.7826)	EN (α = 0.0, λ = 4.9770)	EN (α = 0.0, λ = 2.2051)	EN (α = 0.0, λ = 3.1257)	EN (α = 0.0, λ = 2.7826)	SPLS (K = 2, η = 0.2)
5	EN (α = 0.0, λ = 2.7826)	EN (α = 0.0, λ = 4.4306)	EN (α = 1.0, λ = 0.0210)	SPLS (K = 3, η = 0.4)	EN (α = 0.0, λ = 2.4771)	EN (α = 0.0, λ = 3.9442)
6	EN (α = 1.0, λ = 0.0059)	EN (α = 0.0, λ = 0.0074)	EN (α = 0.1, λ = 0.0210)	EN (α = 1.0, λ = 0.0052)	EN (α = 0.5, λ = 0.0007)	SPLS (K = 2, η = 0.8)
7	EN (α = 0.6, λ = 0.0023)	EN (α = 0.0, λ = 0.0059)	EN (α = 0.1, λ = 0.0004)	EN (α = 0.0, λ = 0.0013)	EN (α = 1.0, λ = 0.0003)	EN (α = 1.0, λ = 0.0023)
8	OLS	EN (α = 0.7, λ = 0.0006)	EN (α = 1.0, λ = 0.0066)	EN (α = 1.0, λ = 0.0005)	SPLS (K = 8, η = 0.9)	SPLS (K = 10, η = 0.0)
9	EN (α = 0.0, λ = 0.0029)	SPLS (K = 3, η = 0.8)	SPLS (K = 5, η = 0.8)	EN (α = 0.0, λ = 0.0167)	SPLS (K = 5, η = 0.0)	EN (α = 0.0, λ = 0.0266)
10	EN (α = 0.9, λ = 0.0041)	EN (α = 0.4, λ = 0.0187)	EN (α = 0.8, λ = 0.0037)	EN (α = 0.0, λ = 0.0534)	EN (α = 0.0, λ = 0.0148)	SPLS (K = 9, η = 0.8)
11	EN (α = 0.1, λ = 0.1072)	EN (α = 0.0, λ = 0.2154)	EN (α = 0.1, λ = 0.1072)	EN (α = 0.4, λ = 0.0534)	EN (α = 0.3, λ = 0.0266)	EN (α = 0.5, λ = 0.0475)
12	SPLS (K = 4, η = 0.6)	SPLS (K = 5, η = 0.6)	EN (α = 0.7, λ = 0.0236)	SPLS (K = 5, η = 0.5)	EN (α = 0.0, λ = 0.1520)	EN (α = 1.0, λ = 0.0148)
13	SPLS (K = 5, η = 0.6)	SPLS (K = 5, η = 0.5)	EN (α = 0.2, λ = 0.0534)	EN (α = 0.0, λ = 0.1353)	EN (α = 0.0, λ = 0.1353)	SPLS (K = 5, η = 0.6)
14	EN (α = 0.0, λ = 0.3430)	EN (α = 0.1, λ = 0.1520)	EN (α = 0.1, λ = 0.0673)	EN (α = 0.0, λ = 0.2154)	EN (α = 0.0, λ = 0.2420)	SPLS (K = 5, η = 0.7)

Table C.2. Best-Performing Machine Learning Methods and Hyperparameter Values for First-Stage Ensemble Models

Ensemble	Realistic	Investigative	Artistic	Social	Enterprising	Conventional
1	SPLS (K = 3, η = 0.0)	EN (α = 0.0, λ = 0.2719)	SPLS (K = 2, η = 0.0)	SPLS (K = 2, η = 0.0)	EN (α = 0.0, λ = 0.3054)	EN (α = 1.0, λ = 0.0955)
2	EN (α = 0.7, λ = 0.0423)	EN (α = 0.0, λ = 0.5462)	EN (α = 0.6, λ = 0.0534)	SPLS (K = 1, η = 0.0)	EN (α = 0.0, λ = 0.4329)	SPLS (K = 1, η = 0.0)
3	EN (α = 0.7, λ = 0.0376)	SPLS (K = 1, η = 0.0)	EN (α = 0.6, λ = 0.0534)	SPLS (K = 1, η = 0.0)	SPLS (K = 1, η = 0.0)	SPLS (K = 1, η = 0.0)
4	EN (α = 0.7, λ = 0.0423)	EN (α = 0.0, λ = 0.5462)	EN (α = 0.6, λ = 0.0534)	EN (α = 0.0, λ = 0.3854)	EN (α = 0.0, λ = 0.3854)	SPLS (K = 1, η = 0.8)
5	EN (α = 0.7, λ = 0.0423)	SPLS (K = 1, η = 0.9)	EN (α = 0.6, λ = 0.0534)	SPLS (K = 1, η = 0.0)	SPLS (K = 1, η = 0.8)	SPLS (K = 1, η = 0.0)
6	EN (α = 0.7, λ = 0.0423)	SPLS (K = 1, η = 0.9)	EN (α = 0.6, λ = 0.0534)	EN (α = 0.0, λ = 0.3854)	EN (α = 0.0, λ = 0.4863)	SPLS (K = 1, η = 0.0)
7	EN (α = 0.7, λ = 0.0376)	SPLS (K = 1, η = 0.0)	EN (α = 0.6, λ = 0.0534)	SPLS (K = 1, η = 0.0)	EN (α = 0.0, λ = 0.5462)	SPLS (K = 1, η = 0.0)
8	EN (α = 0.7, λ = 0.0423)	SPLS (K = 1, η = 0.9)	EN (α = 0.6, λ = 0.0534)	EN (α = 0.0, λ = 0.3854)	EN (α = 0.0, λ = 0.4329)	SPLS (K = 1, η = 0.8)
9	EN (α = 0.7, λ = 0.0423)	SPLS (K = 1, η = 0.9)	EN (α = 0.6, λ = 0.0534)	SPLS (K = 1, η = 0.0)	EN (α = 0.0, λ = 0.4863)	SPLS (K = 1, η = 0.0)
10	SPLS (K = 1, η = 0.0)	SPLS (K = 1, η = 0.0)	SPLS (K = 1, η = 0.8)	EN (α = 0.0, λ = 0.3854)	EN (α = 0.0, λ = 0.6136)	SPLS (K = 1, η = 0.0)
11	EN (α = 0.0, λ = 0.2154)	SPLS (K = 1, η = 0.8)	EN (α = 0.0, λ = 0.3054)	EN (α = 0.0, λ = 0.4329)	EN (α = 0.0, λ = 0.5462)	SPLS (K = 1, η = 0.0)
12	SPLS (K = 1, η = 0.0)	SPLS (K = 1, η = 0.0)	SPLS (K = 1, η = 0.8)	EN (α = 0.0, λ = 0.4863)	EN (α = 0.0, λ = 0.6893)	SPLS (K = 1, η = 0.0)
13	EN (α = 0.0, λ = 0.2154)	EN (α = 0.0, λ = 0.6893)	EN (α = 0.0, λ = 0.3430)	EN (α = 0.0, λ = 0.4863)	EN (α = 0.0, λ = 0.4863)	SPLS (K = 1, η = 0.0)
14	EN (α = 0.0, λ = 0.3430)	EN (α = 0.0, λ = 0.7743)	SPLS (K = 1, η = 0.8)	EN (α = 0.0, λ = 0.4863)	EN (α = 0.0, λ = 0.6136)	SPLS (K = 1, η = 0.0)
15	EN (α = 0.0, λ = 0.2719)	EN (α = 0.0, λ = 0.6893)	EN (α = 0.0, λ = 0.3430)	EN (α = 0.0, λ = 0.4863)	EN (α = 0.0, λ = 0.5462)	SPLS (K = 1, η = 0.0)
16	SPLS (K = 1, η = 0.0)	SPLS (K = 1, η = 0.0)	SPLS (K = 1, η = 0.8)	EN (α = 0.0, λ = 0.5462)	EN (α = 0.0, λ = 0.6893)	SPLS (K = 1, η = 0.0)
17	EN (α = 0.0, λ = 0.2719)	EN (α = 0.0, λ = 0.6893)	EN (α = 0.0, λ = 0.3430)	EN (α = 0.0, λ = 0.4863)	EN (α = 0.0, λ = 0.5462)	SPLS (K = 1, η = 0.0)
18	EN (α = 0.0, λ = 0.3854)	EN (α = 0.0, λ = 0.7743)	EN (α = 0.0, λ = 0.4329)	EN (α = 0.0, λ = 0.5462)	EN (α = 0.0, λ = 0.6136)	SPLS (K = 1, η = 0.0)

Appendix D: SME Interest Rating Materials

RIASEC Familiarization Exercise Instructions

I'll present you with an occupation's ***description*** and ***tasks***. Apply lessons learned during construct training to:

- Identify the **top three** RIASEC construct categories (i.e., the high-point codes) most associated with the occupation.
- Explain ***why*** you chose that RIASEC category.

NOTE: Although you won't be identifying high-point codes as part of the ratings you'll be making, the purpose of this activity is just to get you more familiar with the RIASEC model.

Do:

- Review task statements to understand the actions (verbs) and objects necessary to perform the job.
- Rely on the occupational description and how it relates to work involved under each RIASEC category.
- Focus on the tasks performed.
- Ask yourself, "How descriptive and characteristic is the given Holland work environment of this occupation?"

Don't:

- Rely on your personal experience or stereotype of the occupation and tasks performed.

Rating Instructions and Rating Sheet

Instructions: The O*NET 2023 Vocational Interest Data Collection Master Rating Booklet contains all information necessary to rate 269 occupations on Realistic, Investigative, Artistic, Social, Enterprising, and Conventional (RIASEC) interest categories.

Contents (tabs) of Workbook:

Instructions. Includes instructions for making RIASEC ratings.

Master Ratings. The master ratings tab includes the O*NET occupation code and title. This is the worksheet where raters will document their RIASEC ratings (1-7 scale) for each occupation.

Note, the occupations in each tab are grouped into three sets: (a) the first 10 for use in initial training, (b) the next 50 for use in calibration between initial and follow-up training, and (c) the remaining 209 occupations for rating post-calibration. Within each set, occupations are sorted by O*NET-SOC.

Steps:

1. Review the occupational information associated with an occupation before providing a rating (provided in a separate Excel file).

2. Rate the occupation on each RIASEC category.

    a. **Do:** Review task statements to understand the actions (verbs) and objects necessary to perform the job. Rely on the occupational description and how it relates to work involved under each RIASEC category. Focus on the tasks performed. Ask yourself, “How descriptive and characteristic is the given Holland work environment of this occupation?”

    b. **Don’t:** Rely on your personal experience or stereotype of the occupation and tasks performed.

3. Repeat steps 1-2 until all occupations in workbook are rated.

Example Interest Rating Sheet

	Interest Ratings					
	1 = Not at all characteristic; 4 = Moderately characteristic; 7 = Extremely characteristic					
Occupational Title	**Realistic**	**Investigative**	**Artistic**	**Social**	**Enterprising**	**Conventional**
General and Operations Managers						
Advertising and Promotions Managers						
Security Managers						
Quality Control Systems Managers						
Natural Sciences Managers						
Environmental Compliance Inspectors						
Coroners						
Logistics Engineers						
Accountants and Auditors						
Loan Officers						
Fraud Examiners, Investigators and Analysts						
Health Informatics Specialists						
Data Warehousing Specialists						
Web and Digital Interface Designers						
Penetration Testers						
Data Scientists						

Appendix E: Elastic Net Regression Hyperparameter Values by Residual Model

Table E.1. Elastic Net Regression Hyperparameters by Residual Model

Residual Model	Outer Fold	Realistic	Investigative	Artistic	Social	Enterprising	Conventional
RM1	1	α = 0.0, λ = 3.5112	α = 0.1, λ = 0.8697	α = 0.0, λ = 16.0000	α = 0.1, λ = 0.7743	α = 0.0, λ = 16.0000	α = 0.2, λ = 0.4329
	2	α = 0.0, λ = 3.1257	α = 0.0, λ = 17.0000	α = 0.6, λ = 0.1205	α = 0.0, λ = 13.0000	α = 0.0, λ = 4.4306	α = 0.0, λ = 6.2803
	3	α = 0.0, λ = 6.2803	α = 1.0, λ = 0.0673	α = 1.0, λ = 0.1072	α = 0.0, λ = 16.0000	α = 0.0, λ = 11.0000	α = 0.4, λ = 0.2154
	4	α = 0.0, λ = 4.9770	α = 0.0, λ = 15.0000	α = 0.0, λ = 15.0000	α = 0.0, λ = 12.0000	α = 0.0, λ = 5.5908	α = 0.1, λ = 0.5462
	5	α = 0.0, λ = 7.0548	α = 0.0, λ = 17.0000	α = 1.0, λ = 0.1353	α = 0.0, λ = 13.0000	α = 0.1, λ = 0.9770	α = 0.0, λ = 7.0548
RM2	1	α = 0.2, λ = 0.2154	α = 1.0, λ = 0.0756	α = 0.0, λ = 35.0000	α = 0.1, λ = 0.3854	α = 0.0, λ = 7.9248	α = 0.0, λ = 34.0000
	2	α = 0.2, λ = 0.1520	α = 0.0, λ = 22.0000	α = 0.0, λ = 7.0548	α = 0.0, λ = 19.0000	α = 0.0, λ = 3.1257	α = 0.3, λ = 0.1707
	3	α = 0.1, λ = 0.3054	α = 0.0, λ = 30.0000	α = 0.0, λ = 7.0548	α = 0.0, λ = 11.0000	α = 0.0, λ = 5.5908	α = 0.6, λ = 0.1353
	4	α = 0.0, λ = 3.9442	α = 0.0, λ = 34.0000	α = 0.1, λ = 0.3854	α = 0.6, λ = 0.1353	α = 0.0, λ = 7.0548	α = 0.0, λ = 24.0000
	5	α = 0.0, λ = 1.9630	α = 0.0, λ = 29.0000	α = 0.0, λ = 7.0548	α = 0.0, λ = 12.0000	α = 1.0, λ = 0.0673	α = 0.0, λ = 21.0000
RM3	1	α = 0.3, λ = 0.1707	α = 0.0, λ = 36.0000	α = 0.0, λ = 11.0000	α = 0.0, λ = 4.9770	α = 0.1, λ = 0.6136	α = 0.0, λ = 30.0000
	2	α = 0.0, λ = 2.7826	α = 0.5, λ = 0.1205	α = 0.0, λ = 7.9248	α = 0.2, λ = 0.3430	α = 0.1, λ = 0.4329	α = 0.0, λ = 7.0548
	3	α = 0.8, λ = 0.0850	α = 0.0, λ = 25.0000	α = 0.0, λ = 15.0000	α = 0.0, λ = 8.9022	α = 0.0, λ = 17.0000	α = 0.1, λ = 0.5462
	4	α = 0.0, λ = 4.4306	α = 0.0, λ = 50.0000	α = 0.0, λ = 10.0000	α = 0.0, λ = 8.9022	α = 0.1, λ = 0.6893	α = 0.0, λ = 29.0000
	5	α = 0.0, λ = 5.5908	α = 0.0, λ = 31.0000	α = 0.0, λ = 4.9770	α = 0.1, λ = 0.2719	α = 0.1, λ = 1.0975	α = 0.1, λ = 0.7743

Appendix F: Final RIASEC Model Residuals by O*NET Job Family and Job Zone

Table F.1. Raw Residual Summary by Job Family

Job Family		Raw Residual											
		R		I		A		S		E		C	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Architecture and Engineering	14	.04	.52	.00	.68	.00	.55	-.08	.41	.36	.44	-.04	.37
Arts, Design, Entertainment, Sports, and Media	19	.09	.62	.07	.64	-.62	.72	.07	.42	-.24	.62	.16	.66
Building Grounds Cleaning and Maintenance	1	.00		.67		-.33		-.38		.42		.47	
Business and Financial Operations	22	.14	.76	-.14	.90	.13	.35	.16	.42	.44	.71	-.12	.65
Community and Social Service	2	.23	.02	-.10	.59	.88	.34	.00	.15	.69	.60	-.27	.09
Computer and Mathematical	20	.18	.61	.29	.74	.14	.56	-.14	.61	.02	.49	-.10	.46
Construction and Extraction	8	-.12	.15	.40	.53	.16	.25	.25	.19	-.20	.99	.16	.47
Educational Instruction and Library	22	-.06	.63	-.04	.43	.09	.70	-.27	.41	.05	.56	-.09	.57
Farming, Fishing, and Forestry	4	-.27	.21	-.05	.80	.11	.33	-.40	1.00	.26	.43	-.05	.76
Food Preparation and Serving Related	7	.23	.51	.00	.16	-.34	1.15	.14	.74	.28	.64	.64	.75
Healthcare Practitioners and Technical	36	.09	.59	.27	.72	.10	.37	.08	.78	-.06	.60	-.12	.61
Healthcare Support	3	.43	1.24	.16	.14	.41	.70	.43	.44	.37	.19	-.56	.58
Installation, Maintenance, and Repair	6	-.02	.22	.38	.32	.12	.16	-.05	.31	-.15	.42	.13	.77
Legal	3	.08	.08	.79	.80	-.07	.68	.55	.36	-1.22	.48	-.26	.83
Life, Physical, and Social Science	18	-.11	.60	-.15	.62	.02	.67	-.17	.61	.02	.72	.14	.51
Management	21	-.04	.53	.36	.62	.05	.49	-.04	.51	-.27	.63	-.08	.63
Office and Administrative Support	6	.33	.77	.51	.73	.15	.25	.27	.39	.00	.70	-.25	.63
Personal Care and Service	13	.09	.51	-.13	.45	.07	.61	-.16	.66	.34	.63	.09	.95
Production	11	.24	.50	.47	.53	.50	.92	.03	.26	.13	.37	-.15	.72
Protective Service	12	-.05	.80	.22	.67	.07	.24	.42	.54	.25	.84	-.17	.94
Sales and Related	5	-.23	.45	-.38	1.29	.35	.20	.41	.47	-1.21	.67	.74	.18
Transportation and Material Moving	16	.10	1.11	.13	.44	.04	.15	.03	.55	.11	.85	.10	.83
All O*NET-SOCs in Sample	269	.06	.63	.13	.67	.05	.58	.02	.57	.03	.69	-.01	.66

Note. *n* = number of occupations in job family. *M* = mean raw residual across occupations. *SD* = standard deviation of raw residuals across occupations. Positive mean values indicate that predicted scores are higher than expert scores, and negative mean valued indicate predicted scores are lower than expert scores on average. Mean values are shaded along a green-red color gradient to facilitate interpretation (higher values – indicating overprediction – are shaded red, lower values – indicating underprediction –are shaded green).

Table F.2. Absolute Residual Summary by Job Family

Job Family		Absolute Residual											
		R		I		A		S		E		C	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Architecture and Engineering	14	.39	.33	.55	.37	.42	.33	.30	.28	.42	.39	.31	.19
Arts, Design, Entertainment, Sports, and Media	19	.52	.32	.55	.32	.79	.52	.32	.28	.54	.37	.53	.40
Building Grounds Cleaning and Maintenance	1	.00		.67		.33		.38		.42		.47	
Business and Financial Operations	22	.55	.54	.69	.58	.22	.30	.35	.26	.71	.42	.45	.47
Community and Social Service	2	.23	.02	.42	.14	.88	.34	.11	.01	.69	.60	.27	.09
Computer and Mathematical	20	.50	.37	.47	.63	.43	.38	.44	.44	.37	.31	.39	.26
Construction and Extraction	8	.14	.13	.46	.47	.17	.23	.27	.16	.78	.57	.39	.28
Educational Instruction and Library	22	.44	.45	.35	.25	.59	.37	.38	.31	.40	.38	.44	.36
Farming, Fishing, and Forestry	4	.27	.21	.60	.41	.28	.14	.75	.68	.42	.20	.56	.40
Food Preparation and Serving Related	7	.38	.38	.10	.12	.88	.74	.59	.40	.60	.28	.69	.70
Healthcare Practitioners and Technical	36	.47	.35	.62	.45	.28	.26	.65	.42	.42	.42	.47	.40
Healthcare Support	3	.99	.59	.16	.14	.41	.70	.43	.44	.37	.19	.56	.58
Installation, Maintenance, and Repair	6	.13	.16	.39	.30	.12	.16	.25	.15	.31	.30	.63	.36
Legal	3	.08	.08	.86	.70	.53	.20	.55	.36	1.22	.48	.55	.58
Life, Physical, and Social Science	18	.37	.48	.45	.44	.49	.44	.49	.39	.47	.54	.41	.33
Management	21	.36	.38	.57	.42	.38	.31	.40	.30	.49	.47	.51	.36
Office and Administrative Support	6	.67	.42	.75	.42	.15	.25	.37	.27	.46	.49	.35	.57
Personal Care and Service	13	.41	.30	.32	.33	.50	.33	.52	.41	.59	.38	.66	.67
Production	11	.39	.38	.61	.35	.61	.84	.24	.10	.29	.26	.56	.44
Protective Service	12	.58	.53	.50	.48	.13	.22	.54	.40	.66	.55	.80	.46
Sales and Related	5	.32	.38	.92	.88	.35	.20	.55	.21	1.21	.67	.74	.18
Transportation and Material Moving	16	.79	.76	.34	.29	.08	.14	.41	.36	.63	.56	.60	.55
All O*NET-SOCs in Sample	269	.46	.44	.51	.45	.40	.42	.44	.36	.52	.45	.50	.42

Note. *n* = number of occupations in job family. *M* = mean absolute residual across occupations. *SD* = standard deviation of absolute residuals across occupations. Mean values are shaded along a green-red color gradient to facilitate interpretation (higher values – indicating larger deviation between predicted and expert ratings – are shaded red, lower values – smaller deviation between predicted and expert ratings –are shaded green).

Table F.3. Raw Residual Summary by Job Zone

Job Zone (Degree of Preparation Required)		Raw Residual											
		R		I		A		S		E		C	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1 (Little or none)	6	-.15	.15	.29	.38	-.05	.31	-.07	.52	.48	.47	.50	.87
2 (Some)	53	.10	.81	.23	.49	.17	.49	.06	.56	.13	.79	.18	.75
3 (Medium)	76	.07	.66	.15	.63	-.06	.62	.07	.68	.05	.66	-.07	.65
4 (Considerable)	86	.04	.55	.06	.81	.05	.57	.02	.42	.06	.64	-.08	.58
5 (Extensive)	48	.04	.54	.11	.64	.12	.62	-.08	.61	-.21	.69	-.05	.60
All O*NET-SOCs in Sample	269	.06	.63	.13	.67	.05	.58	.02	.57	.03	.69	-.01	.66

Note. *n* = number of occupations in job zone. *M* = mean raw residual across occupations. *SD* = standard deviation of raw residuals across occupations. Positive mean values indicate that predicted scores are higher than expert scores, and negative mean valued indicate predicted scores are lower than expert scores on average. Mean values are shaded along a green-red color gradient to facilitate interpretation (higher values – indicating overprediction – are shaded red, lower values – indicating underprediction –are shaded green).

Table F.4. Absolute Residual Summary by Job Zone

Job Zone (Degree of Preparation Required)		Absolute Residual											
		R		I		A		S		E		C	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1 (Little or none)	6	.15	.15	.34	.32	.22	.20	.41	.26	.58	.29	.66	.73
2 (Some)	53	.56	.59	.40	.36	.30	.42	.43	.35	.59	.53	.61	.47
3 (Medium)	76	.48	.45	.52	.39	.37	.50	.52	.43	.49	.43	.47	.45
4 (Considerable)	86	.41	.36	.59	.56	.43	.37	.33	.26	.51	.39	.45	.38
5 (Extensive)	48	.42	.33	.52	.39	.52	.36	.50	.35	.51	.50	.50	.33
All O*NET-SOCs in Sample	269	.46	.44	.51	.45	.40	.42	.44	.36	.52	.45	.50	.42

Note. *n* = number of occupations in job zone. *M* = mean absolute residual across occupations. *SD* = standard deviation of absolute residuals across occupations. Mean values are shaded along a green-red color gradient to facilitate interpretation (higher values – indicating larger deviation between predicted and expert ratings – are shaded red, lower values – smaller deviation between predicted and expert ratings –are shaded green).