HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2025 No. 130

Using a Hybrid Artificial Intelligence-Expert Method to Develop Work Style Ratings for the O*NET Database

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 Jiayi Liu, HumRRO Felix Wu, HumRRO Maura Burke, HumRRO Phil Lewis, National Center for O*NET Development	**Date:**	December 9, 2025

Using a Hybrid Artificial Intelligence-Expert Method to Develop Work Style Ratings for the O*NET Database

Table of Contents

Table of Contents (Continued)

List of Tables

List of Figures

Using a Hybrid Artificial Intelligence-Expert Method to Develop Work Style Ratings for the O*NET Database

Introduction

The Occupational Information Network (O*NET) is a comprehensive system developed by the U.S. Department of Labor that provides information on over 900 occupations that cover over 55,000 jobs within the U.S. economy. This information is maintained in a comprehensive database (National Center for O*NET Development, 2025). 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 to identify and maintain current information on the characteristics of workers and occupations. This report summarizes efforts to develop, evaluate, and implement an efficient, accurate method for populating Work Style ratings in O*NET that reflects a blend of generative Artificial Intelligence (AI) and human expert judgment. This work represents the final phase of a multi-year effort by the Center to revisit and update the Work Styles content and data within O*NET. The first phase focused on reviewing and updating the content of the Work Styles taxonomy itself (see Putka et al., 2024). As part of the current effort, the focus shifted to developing, evaluating, and implementing an efficient and accurate method for populating Work Style ratings for all active, data-level occupations in O*NET.

Work Styles constitute one type of worker attribute in the O*NET Content Model. Before the first phase of work described above, Work Styles were formally defined within O*NET as "personal characteristics that are work- and job-related" (Borman et al., 1999, p. 213); they are essentially non-clinical personality traits that are relevant to jobs, occupations, and the world of work more broadly. O*NET's Work Styles were originally developed by Borman and colleagues (1995) based on a review of relevant literature (e.g., Industrial-Organizational [I-O] psychology, management, and personality), with a focus on content documenting extant taxonomies, structure, and internal relations, and the job- and work-relatedness of personality traits. Based on this review, they derived seven broad, higher-order Work Style dimensions and 17 more specific, lower-order Work Style dimensions that function as constituent elements of the broader dimensions. Work Styles were initially viewed as being of use in aiding personnel selection, employment counseling, and self-directed job search (Borman et al., 1995), and they remain an essential part of the O*NET Program's support for educational planning, career exploration, career guidance, job search, and organizational placement.

Revisiting Work Styles in the O*NET Content Model

Until the first phase of work that revisited and updated the Work Styles taxonomy mentioned above, O*NET's Work Styles had changed very little in the 30 years since their formulation. In light of the enormous growth in scholarship related to personality in the nearly 30 years since Borman and colleagues (1995) derived O*NET's Work Styles, in early 2024, the Center decided to revisit the Work Styles portion of the O*NET Content Model and update it, considering contemporary developments. This first phase of work involved (a) conducting a review and analysis of relevant literature on personality structure and personality traits relevant to work to inform revisions to the Work Styles, (b) conducting a review of strategies for scaling Work Styles for job analysis applications to inform revisions to potential rating scales for Work Styles in O*NET, (c) drafting updated Work Style content (i.e., higher order and lower order dimensions and descriptions) for the O*NET Content Model, and (d) linking the draft Work Styles to

O*NET's Generalized Work Activities (GWAs) and Work Contexts (WCs) and using the results of that mapping to help finalize an updated set of Work Style dimensions.

The results of this "Phase 1" work led to a *provisional*, updated Work Styles taxonomy consisting of 21 lower-order Work Style dimensions (i.e., Work Styles for which ratings will be populated in the O*NET database) and seven higher-order Work Style dimensions that reflect related groupings of the lower-order Work Styles. The taxonomy was viewed as "provisional," given the possibility that it would be revised as a result of analyses conducted as part of the current, final phase of the Work Styles updating effort. Within the provisional taxonomy, Work Styles are defined as "personality tendencies exhibited at work that can affect how well someone performs a job". Table 1 provides the complete set of Work Style dimensions resulting from the first phase of work. Details on the development of these Work Styles are provided in Putka et al. (2024).

Table 1. Provisional Work Styles from Phase 1

Element Name	Description
Conscientiousness	A tendency to exhibit achievement-oriented behavior, dependability, attention to detail, self-confidence, and cautiousness at work.
Achievement Orientation	A tendency to establish and maintain personally challenging work-related goals, set high work-related standards, and exert high effort toward meeting those goals and standards.
Attention to Detail	A tendency to be detail-oriented, organized, and thorough in completing work.
Cautiousness	A tendency to be careful, deliberate, and risk-avoidant when making work-related decisions or doing work.
Dependability	A tendency to be reliable, responsible, and consistent in meeting work-related obligations.
Self-Confidence	A tendency to believe in one's work-related capabilities and ability to control one's work-related outcomes.
Extraversion	A tendency to exhibit leadership and socially oriented behavior at work.
Leadership Orientation	A tendency to lead, take charge, offer opinions, and provide direction at work.
Social Orientation	A tendency to seek out, enjoy, and be energized by social interaction at work.
Agreeableness	A tendency to exhibit cooperative and empathetic behavior towards others at work.
Cooperation	A tendency to be pleasant, helpful, and willing to assist others at work.
Empathy	A tendency to show concern for others and be sensitive to others' needs and feelings at work.
Honesty-Humility	A tendency to exhibit sincerity, humility, and integrity at work.
Humility	A tendency to be modest and humble when interacting with others at work.
Integrity	A tendency to be honest and ethical at work.
Sincerity	A tendency to be genuine and sincere in interactions with others at work, without concern for personal gain or self-interest.

Table 1. (Continued)

Element Name	Description
Emotional Stability	A tendency to exhibit self-control and tolerate stress at work.
Self-Control	A tendency to remain calm and composed and to manage emotions effectively in response to criticism or difficult situations at work.
Stress Tolerance	A tendency to cope and function effectively in stressful situations at work.
Openness	A tendency to exhibit adaptability, flexibility, intellectual curiosity, and innovation at work.
Adaptability	A tendency to be open to and comfortable with change, new experiences, or ideas at work.
Innovation	A tendency to be inventive, to be imaginative, and to adopt new perspectives on ways to accomplish work.
Intellectual Curiosity	A tendency to seek out and acquire new work-related knowledge and obtain a deep understanding of work-related subjects.
Tolerance for Ambiguity	A tendency to be comfortable with ambiguity and uncertainty at work.
Compound Dimensions	Work style dimensions that reflect a combination of elements from two or more broad personality domains.
Initiative	A tendency to be proactive and take on extra responsibilities and tasks that may fall outside of one's required work role.
Optimism	A tendency to exhibit a positive attitude and positive emotions at work, even under difficult circumstances.
Perseverance	A tendency to exhibit determination and resolve to perform or complete tasks in the face of difficult circumstances or obstacles at work.

Note. Lower-order Work Styles (indented) are grouped under higher-order Work Style dimensions.

The Current Project

The objective of the current effort was to develop, evaluate, and implement an efficient, accurate method for populating Work Style ratings for all active, data-level occupations in O*NET. Our emphasis was on identifying a method that could easily be implemented as new occupations emerge or as occupations change (e.g., changes to occupation descriptions or tasks), and that would minimize or eliminate the need for job incumbent, occupational expert (OE) or occupational analyst data collections to keep Work Style ratings up to date in O*NET. To accomplish this objective, we engaged in the following seven steps, which we document in the current report:

- Step 1: Identifying Occupations for Expert and Analyst Data Collections[1]
- Step 2: Collecting Expert and Analyst Work Style Ratings
- Step 3: Evaluation of Expert Ratings as the Gold Standard
- Step 4: Exploration of Alternative Rating Methods

[1] Though we aimed to eliminate the need for expert or analyst Work Style data collections in the future, for purpose of this effort, we gathered ratings from experts on personality in the workplace for a sample of occupations for evaluating alternative rating methods, and also gathered occupational analyst-based ratings as a “back up” strategy in the event we were unable to develop a generative AI-based rating method (or other empirical rating method) that performed effectively.

- Step 5: Evaluation of LLM Ratings Against Expert Ratings
- Step 6: Summary of Draft Final Ratings Across All Data-Level Occupations
- Step 7: Finalizing Work Style Ratings for the O*NET Database

We provide details on each step in the sections that follow. Lastly, we conclude this report with recommendations for populating and updating Work Style ratings in future versions of the O*NET database.

Step 1: Identifying Occupations for Expert and Analyst Data Collections

As part of the current effort, we gathered Work Style ratings from O*NET occupation analysts and a group of Work Style experts for a subset of the 923 data-level O*NET-SOCs (occupations) that were active as of O*NET 28.3 to aid in the development and evaluation effort. We gathered expert ratings to provide a "gold" standard against which we evaluated all other rating methods considered and to aid in calibrating ratings from generative AI-based methods. We gathered analyst ratings to provide a traditional point of comparison for generative AI-based methods, and to serve as a potential alternative method should the generative AI-based and other empirical methods we considered fail to sufficiently mirror expert ratings.

Our intent here was not to gather human ratings for all 923 data-level occupations that were active at the time occupations were identified for the data collection (i.e., July 2024, O*NET 28.3), but rather a representative subset of occupations stratified by job zone and job family. [2] We sampled 125 occupations for inclusion in the data collections.[3] Tables 1.1 and 1.2 provide a comparison of the sample of 125 occupations relative to the full set of 923 data-level occupations in terms of job zone and job family representation. Comparison of the sample of 125 occupations relative to the full set of 923 data-level occupations revealed comparable coverage of all O*NET job zones and job families.

Table 1.1. Representativeness of Occupations Selected for Inclusion in the Expert and Analyst Data Collections with Respect to O*NET Job Zone

Job Zone	All O*NET 28.3 Data-Level Occupations			O*NET 28.3 Data Level Occupations Selected for Inclusion in Data Collections			
	n	%		*n*	%		Δ%
1: Little or no preparation needed	31	3.4		8	6.4		3.0
2: Some preparation needed	293	31.7		38	30.4		-1.3
3: Medium preparation needed	220	23.8		27	21.6		-2.2
4: Considerable preparation needed	224	24.3		30	24.0		-0.3
5: Extensive preparation needed	155	16.8		22	17.6		0.8
Total	923	100		125	100		0

[2] 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).
[3] The sample size of 125 attempted to strike a balance between the level of effort required by analysts and experts to rate the occupations and obtaining a sufficient sample size that would allow for estimation of correlations among Work Style rating methods (i.e., expert, analyst, generative AI, or other empirical alternatives).

Note. Δ% = The percent of O*NET 28.3 data-level occupations selected for inclusion in the Work Styles data collection minus the percent of all O*NET 28.3 data-level occupations in the given job zone.

Table 1.2. Representativeness of Occupations Selected for Inclusion in the Expert and Analyst Data Collections with Respect to Job Family

Job Family	All O*NET 28.3 Data-Level Occupations			O*NET 28.3 Data Level Occupations Selected for Inclusion in Data Collections			
	n	%		*n*	%		Δ%
Architecture and Engineering	56	6.1		8	6.4		0.3
Arts, Design, Entertainment, Sports, and Media	40	4.3		5	4.0		-0.3
Building and Grounds Cleaning and Maintenance	8	0.9		3	2.4		1.5
Business and Financial Operations	48	5.2		8	6.4		1.2
Community and Social Service	14	1.5		2	1.6		0.1
Computer and Mathematical	36	3.9		5	4.0		0.1
Construction and Extraction	61	6.6		7	5.6		-1.0
Educational Instruction and Library	62	6.7		7	5.6		-1.1
Farming, Fishing, and Forestry	12	1.3		3	2.4		1.1
Food Preparation and Serving Related	16	1.7		3	2.4		0.7
Healthcare Practitioners and Technical	89	9.6		10	8.0		-1.6
Healthcare Support	19	2.1		2	1.6		-0.5
Installation, Maintenance, and Repair	50	5.4		6	4.8		-0.6
Legal	7	0.8		3	2.4		1.6
Life, Physical, and Social Science	60	6.5		7	5.6		-0.9
Management	56	6.1		8	6.4		0.3
Office and Administrative Support	51	5.5		6	4.8		-0.7
Personal Care and Service	31	3.4		5	4.0		0.6
Production	107	11.6		12	9.6		-2.0
Protective Service	26	2.8		3	2.4		-0.4
Sales and Related	22	2.4		4	3.2		0.8
Transportation and Material Moving	52	5.6		8	6.4		0.8
Total	923	100		125	100		0

Note. Δ% = % of O*NET 28.3 data-level occupations selected for inclusion in the Work Styles data collection minus % of all O*NET 28.3 data-level occupations in the given job zone.

Step 2: Collecting Expert and Analyst Work Style Ratings

HumRRO recruited trained O*NET occupation analysts, as well as academic researchers with expertise in the study of personality in the context of work, to provide Work Style ratings for the 125 occupations identified in Step 1.

Overview of Rater Recruitment and Training

We recruited eight O*NET analysts with experience in the O*NET Data Collection Program. These analysts were selected based on their proven ability to provide high quality ratings in a timely fashion. Additionally, we recruited five academic researchers who professionally study and publish on personality in the workplace and have deep familiarity with O*NET and occupational analysis. Those individuals were Dr. Ann Marie Ryan of Michigan State University's Department of Psychology, Dr. Christopher Nye of Michigan State University's Department of Psychology, Dr. Chad Van Iddekinge of the University of Iowa's Tippie College of Business, Dr. Deniz Ones of the University of Minnesota's Department of Psychology, and Dr. Fred Oswald of Rice University's Department of Psychological Sciences.

Rater Training

HumRRO structured rater training strategically such that the expert session occurred first and the analyst training occurred shortly thereafter. The rationale for doing so was twofold. First, we believed experts might have feedback on the rating process given their subject matter expertise, and we wanted to identify ways to improve it (if possible), considering their perspectives. We wanted to address any potential concerns at the outset of training and then reflect any needed changes in guidance we subsequently provided to O*NET analysts. Second, we believed that experts could provide insight into the rating experience (e.g., thoughts or mental processes involved in approaching the task) that we could relay to analysts as guidance. The substantive nature of training presented to each group was largely comparable where HumRRO provided: (a) an overview and purpose of the rating task, (b) a background and brief history of O*NET Work Styles, (c) an overview of the new Work Styles, and (d) a summary of the ratings process, materials, and guidance for provisioning ratings. Both groups also participated in a calibration exercise to establish a common frame of reference for making ratings independently.

As noted above, during the expert rating, several insights emerged during the calibration discussions that helped lead to refinements in guidance given to both experts and analysts. The insights focused mainly on what O*NET occupation-level information should be considered and when during the rating process. A summary of those conversations is distilled below:

- Initially, raters were instructed to review task importance information for the occupation to get a sense for the tasks performed. This included reviewing both core and supplemental tasks for an occupation. However, experts shared that this was a lot of information to process and was sometimes distracting (supplemental tasks). As a result, we updated our guidance for raters to focus solely on core tasks (ignoring supplemental). For occupations that lacked task importance data, raters were asked to consider all associated tasks for those occupations.
- Experts shared that data on linkages between Work Styles and O*NET Generalized Work Activities (GWAs) and between Work Styles and O*NET Work Contexts (WC) that were provided as stimulus materials for the Work Style rating task could sometimes change or contradict the rating an expert was initially thinking of for a given Work Style-

occupation combination, as opposed to corroborating it. As such, we discussed with experts how and when linkage information should be used to inform ratings. We came to an agreement that raters should first make an initial Work Style rating *after* reviewing the occupational description, task, and skill information, but *prior* to reviewing the Work Style-GWA/WC linkage noted above. Once the rater makes the initial rating, they would then consider whether that linkage data should affect their preliminary rating.

After concluding the first expert training session, HumRRO updated the Rating Instructions and Scale information document (see Appendix A) for both groups of raters[4]. This document contained updated steps on how to approach the rating task. These materials were provided to raters to analysts in advance of their training session, which occurred a week or so after the experts. Like the expert training, we provided raters with an overview of the new Work Styles and their definitions, and time was spent discussing the process for making ratings and the rating scale.

Raters were instructed to make their ratings on a 7-point "impact" rating scale (see Figure 2.1) that asks raters whether high standing on the given trait is beneficial or detrimental to performance in the occupation" (see Figure 2.1). In other words, to what extent does the Work Style have a positive or negative *impact* on performance in the occupation? This rating scale was developed as part of Phase 1 of the Work Styles revision effort noted earlier (Putka et al., 2024) based on a review of the personality-oriented job analysis literature.

Figure 2.1. Work Style Impact Rating Scale

-3	-2	-1	0	+1	+2	+3
Very detrimental to job performance in this occupation	Detrimental to job performance in this occupation	Somewhat detrimental to job performance in this occupation	Little or no impact on job performance in this occupation	Somewhat beneficial to job performance in this occupation	Beneficial to job performance in this occupation	Very beneficial to job performance in this occupation

To simplify the use of this rating scale, raters were instructed to think about their ratings as a two-part judgment. First, we instructed raters to ask themselves whether the Work Style was even "relevant" to performance in the given occupation. If it was not, then they were instructed to assign a rating of "0". If the Work Style was relevant, then we instructed raters to ask themselves whether it was "detrimental or beneficial" in that occupation. If they deemed it beneficial (e.g., facilitating) to performance in the occupation, then they would assign a rating of between +1 and +3 depending on how beneficial they believed it was to performance in that occupation. If they deemed it detrimental (e.g., inhibiting) to performance in the occupation, then they would assign a rating of between -1 and -3 depending on how detrimental they believed it was to performance in that occupation.

[4] The guidance given to analysts and raters appearing in Appendix A was the same. The only exception was content regarding the timeline to submit materials, which differed for experts and analysts. This difference in administrative details was omitted from Appendix A for the sake of parsimony.

To clarify the first part of the judgment process, we spent time discussing the concepts of trait activation and relevance. Specifically, the observation that work tasks, situations, and environments differ in the extent to which they offer the opportunity for a personality trait to be expressed or differ in the extent to which they activate or cue expression of a given personality trait. We provided examples such as "Social Orientation" being activated in social situations and tasks involving interacting with others, but *not* activated when one is "meditating in an empty room", since the task doesn't involve interacting with others. The latter reflects a situation where the Work Style is unlikely to be activated and, in turn, is not relevant (addressing the first part of the judgment process).

Following the introduction of the scale, we illustrated Work Styles that were beneficial, detrimental, or irrelevant to various O*NET GWAs and Work Context (based on the linkage data described above, which were gathered during the Phase 1 effort). This was designed to help anchor the discussion of trait activation and trait relevance with real examples grounded in Work Style-GWA/WC linkage data.

Overview of Rating Process

In advance of their training sessions, HumRRO distributed rating materials to analysts and experts. Note that the materials were identical across rater groups, as experts and analysts rated the same occupations. The materials consisted of four Excel files: (a) stimulus materials (i.e., information on the occupations and Work Styles to be rated) for two sets of occupations (i.e., 25 occupations for calibration, and the remaining 100 occupations to be rated) and (b) rating booklets for the two sets of occupations, where raters entered their ratings each Work Style. The stimulus workbooks contained nine tabs worth of information, all drawn from the updated O*NET 29.0 database; the tabs included the following information:

- **Tab 1:** O*NET-SOC 2019 code, occupation name, occupation job zone, and occupation description.
- **Tab 2**: Occupation "core" tasks (i.e., tasks that have relevance ratings ≥ 67% and an importance rating of ≥ 3.0 [on a 1-5 importance scale] for a given occupation), task type (core, supplemental, or n/a), and task importance ratings.[5] For occupations that had not yet had their task ratings populated in O*NET 29.0, all tasks associated with that occupation were listed. For each occupation, tasks were sorted in descending order of importance. If tasks had no importance rating for an occupation, they were listed in alphabetical order.
- **Tab 3:** Occupation skills and skill importance ratings. Skills presented for each occupation were limited to those that had importance ratings of ≥ 3.0 [on a 1-5 importance scale] for a given occupation. If an occupation had more than 10 skills with importance ratings ≥ 3.0, only the top 10 skills were listed. For each occupation, skills were sorted in descending order of importance.
- **Tab 4:** GWAs and linkage ratings between GWAs and each Work Style. GWAs were limited to roughly the top 7 most important GWAs for each occupation. For each occupation, GWAs were sorted in a manner that highlighted the GWAs that best

[5] The expert stimulus materials included both core and supplemental tasks when task importance ratings were available for an occupation, but for the analyst materials, only core tasks were included when task importance ratings were available for an occupation. As noted above, experts suggested focusing only on the core task when available, so the supplemental tasks were removed from the analyst stimulus materials.

differentiated that occupation from other occupations based on the inverse of the frequency with which the GWA was deemed important across occupations.[6] Also included in this tab are the linkage ratings between each GWA and the 21 Work Styles for those GWA-Work Style combinations that were linked by Subject Matter Experts (SMEs) in the first phase of the Work Styles revision effort.

- **Tab 5:** Work Contexts and linkage ratings between Work Contexts and each Work Style. Work Contexts were limited to roughly the top 7 most important Work Contexts for each occupation. For each occupation, Work Contexts were sorted in a manner that highlighted the Work Contexts that best differentiated that occupation from other occupations based on the inverse frequency with which the Work Context was deemed important across occupations. Also included in this tab are the linkage ratings between each Work Context and the 21 Work Styles for those Work Context-Work Style combinations that were linked by SMEs in the first phase of the Work Styles revision effort.
- **Tab 6**: Work Style names, definitions, and higher-order dimensions.
- **Tab 7**: Definitions of O*NET skills included in Tab 3.
- **Tab 8:** Definitions of O*NET GWAs listed in Tab 4.
- **Tab 9:** Definitions O*NET Work Contexts listed in Tab 5.

In addition to the stimulus materials and rating booklet, HumRRO provided raters with a copy of the training slides as well as a Word document with instructions for making ratings (see Appendix A).

The rating process consisted of three phases:

1) **Initial calibration and group discussion**: Given the sheer number of Work Styles to rate per occupation (i.e., 21), we conducted an initial calibration based on three occupations. These occupations were intentionally selected to reflect different job families with perceived distinct Work Style profiles, and therefore, introduce variation in the Work Styles relevant to a given occupation, fostering a richer calibration discussion. Raters were provided time to independently make their ratings and separately email them back to the trainer. HumRRO then compiled the ratings and flagged Work Style-occupation combinations where clear disagreements were observed (e.g., raters choosing different sides of the beneficial/detrimental scale). The remainder of the training was spent discussing these ratings.

   At the end of initial training, raters were assigned to independently rate an additional 22 occupations asynchronously (bringing the total number of occupations used for calibration to 25). They were also instructed to reconsider and finalize their ratings for the first three occupations, taking into account any shifts in judgment that had occurred during the calibration discussion. Raters were instructed to complete and submit their ratings before a scheduled recalibration session.

6 This sorting strategy was designed to highlight GWAs that were important to the given occupation, yet that were least frequently identified as important across all occupations, to help better differentiate the given occupations in terms of the important GWAs and, in turn, Work Styles linked to those GWAs.

- **Follow-up calibration and group discussion**: In preparation for a follow-up calibration discussion, HumRRO reviewed expert and analyst Work Style ratings for all 25 occupations in the full calibration set, and flagged Work Style-occupation combinations where there were high levels of disagreement within each group (i.e., among experts, and among analysts, respectively), and prioritized discussing those Work Style-occupation combinations during a follow-up calibration meeting with experts and analysts (separately). After having a general discussion about how the independent rating exercise went, we discussed the Work Style-occupations flagged for disagreement and worked with expert and analyst rater groups to further refine a shared rating policy.

2) **Independent rating of remaining occupations**: Upon completion of the calibration effort above, raters independently rated the remaining 100 occupations.

Rating Data Review and Cleaning

Once all final ratings were obtained from experts, we conducted several checks to ensure that the information captured in the individual rating workbooks was both accurate and complete. This involved ensuring that each Work Style-occupation combination had a rating, and following up with individuals when ratings were incomplete. These checks also involved ensuring ratings were within the expected scale range (i.e., -3 to +3) for every Work Style-occupation combination.

Given our plans to use the expert ratings as the gold standard, we took steps to reduce the impact of outlying ratings before proceeding with further analyses with the expert ratings. Specifically, for each of the 2,625 Work Style-occupation combinations rated (21 Work Styles x 125 occupations), we flagged an expert rater as a potential outlier from the group if their rating differed from the mean rating of the rest of the group by two or more scale points. We then treated that rater's rating as missing for the purpose of computing the final mean expert rating for the given Work Style-occupation combination. Of the 2,625 Work Style-occupation combinations rated, 425 (16.2%) had one rater flagged as an outlier, 31 (1.2%) had two raters flagged as outliers, and 3 (0.1%) had three raters flagged as outliers. For the 425 Work Style-occupation combinations where only one expert rater was flagged as an outlier, we used the mean rating of the other four (non-outlying) raters as the rating for that combination. For the 34 Work Style-occupation combinations where more than one expert rater was flagged as an outlier, we used the mean rating of all five raters as the rating for that combination, as the outliers tended to offset each other, so eliminating them as a set did little to change the final rating for the combination.

To treat analyst ratings similarly to expert ratings, we also took steps to reduce the impact of outlying ratings before proceeding with further analyses using the analyst ratings. Specifically, for each of the 2,625 Work Style-occupation combinations rated, we flagged an analyst rater as a potential outlier from the group if their rating differed from the mean rating of the rest of the group by two or more scale points. We then treated that rater's rating as missing for the purpose of computing the final mean analyst rating for the given Work Style-occupation Style combination. Of the 2,625 occupation-Work Style combinations rated, 127 (4.8%) had one rater flagged as an outlier, 11 (0.4%) had two raters flagged as outliers, and 1 (< 0.1%) had four raters flagged as outliers. For the 127 Work Style-occupation combinations where only one analyst rater was flagged as an outlier, we used the mean rating of the other seven (non-outlying) raters as the rating for that combination. For the 12 Work Style-occupation combinations where more than one analyst rater was flagged as an outlier, we used the mean

rating of all eight raters as the rating for that combination, as the outliers again tended to offset each other, so eliminating them as a set did little to change the final rating for the combination.

Step 3: Evaluation of Expert Ratings as the Gold Standard

Given our intent was to use expert ratings as the gold standard against which other rating methods were evaluated, this step focused on evaluating their psychometric quality. Specifically, we aimed to evaluate whether expert ratings for each Work Style exhibited (a) variance across occupations, (b) sufficient levels of interrater reliability and agreement, and (c) evidence of being empirically distinct (i.e., not perfectly, or nearly perfectly correlated). All of the analyses that follow were conducted on the cleaned expert ratings resulting from Step 2.

Basic Descriptives

We first calculated descriptive statistics for the cleaned expert Work Style ratings. Table 3.1 presents descriptive summaries for each Work Style dimension. Work Styles in Table 3.1 are sorted in descending order of their mean rating across occupations and color-coded to facilitate interpretation. Dependability and Attention to Detail exhibited the highest mean ratings, whereas Tolerance for Ambiguity and Humility exhibited the lowest mean ratings. Additionally, all Work Styles appeared to exhibit notable variance (with the exception of Dependability), although the variation was primarily confined to the positive end of the scale. In general, these findings were not surprising, in that it is hard to conceive of an occupation where being highly Dependable is not beneficial to job performance. Most of the Work Styles are generally "positive" attributes, so we would expect work situations where having high levels of the attribute to be detrimental to performance to be rare, and isolated to Work Styles without a clear nexus to Conscientiousness.

Table 3.1. Descriptive Statistics for Work Styles across Occupations

Work Style	*M*	*SD*	*Min*	*Max*	Percentiles				
					5	25	50	75	95
Dependability	2.41	0.35	1.00	3.00	2.00	2.20	2.40	2.60	3.00
Attention To Detail	2.27	0.58	1.00	3.00	1.20	2.00	2.40	2.80	3.00
Integrity	1.68	0.71	0.25	3.00	0.50	1.00	2.00	2.20	2.80
Cautiousness	1.67	0.82	-0.80	3.00	0.12	1.00	2.00	2.25	2.80
Perseverance	1.47	0.56	0.20	3.00	0.60	1.00	1.40	1.80	2.40
Achievement Orientation	1.44	0.66	0.20	3.00	0.40	1.00	1.40	1.80	2.60
Stress Tolerance	1.41	0.64	0.20	3.00	0.60	1.00	1.20	2.00	2.74
Self-Control	1.34	0.72	0.00	3.00	0.40	0.80	1.20	1.80	2.80
Cooperation	1.32	0.79	-0.20	3.00	0.20	0.80	1.20	1.80	2.71
Adaptability	1.19	0.72	-0.20	2.50	0.06	0.60	1.20	1.80	2.36
Self-Confidence	1.17	0.64	0.00	2.80	0.26	0.60	1.00	1.60	2.40
Intellectual Curiosity	1.10	1.00	-0.60	3.00	-0.25	0.20	1.00	2.00	2.80
Initiative	1.03	0.65	-0.50	2.40	0.00	0.50	1.00	1.60	2.14
Social Orientation	0.89	0.97	-0.80	3.00	-0.60	0.00	1.00	1.60	2.47
Sincerity	0.82	0.67	-0.25	3.00	0.00	0.40	0.60	1.00	2.20

Table 3.1. (Continued)

Work Style	*M*	*SD*	*Min*	*Max*	Percentiles				
					5	25	50	75	95
Empathy	0.72	0.93	-0.60	3.00	-0.20	0.00	0.40	1.40	2.60
Innovation	0.71	1.01	-1.00	2.80	-0.74	-0.10	0.60	1.45	2.50
Optimism	0.70	0.80	-0.20	2.60	-0.20	0.00	0.40	1.20	2.25
Leadership Orientation	0.66	1.02	-0.80	3.00	-0.60	-0.20	0.40	1.30	2.57
Tolerance For Ambiguity	0.64	0.87	-1.25	2.60	-0.80	0.00	0.80	1.20	2.00
Humility	0.31	0.62	-2.00	1.80	-0.60	0.00	0.20	0.78	1.50

Note. *N* = 125 (unit of analysis is occupation). Ratings were made on a 7-point scale ranging from -3 (very detrimental to job performance) to +3 (very beneficial to job performance), with a 0-point corresponding to "little or no impact on job performance". Work Styles are sorted in descending order of their mean importance across the 125 occupations sampled.

Interrater Reliability and Agreement

Next, we examined the interrater reliability (i.e., consistency) and absolute agreement (i.e., interchangeability) of the expert Work Style ratings. Table 3.2 shows the interrater reliability and agreement estimates for each Work Style. Two pairs of Intraclass Correlation Coefficient-based (ICC-based) reliability and agreement coefficients are provided (McGraw & Wong, 1996). ICC(C,1) reflects the estimated reliability of a single-rater's rating for the given Work Style dimension. 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 Work Style dimension (across *k* raters). 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 experts, the number of raters (*k*) equals five. Similarly, ICC(A,1) reflects the estimated absolute agreement of a single-rater's rating for the given Work Style dimension (among experts). ICC(A,*k*) reflects the estimated absolute agreement of the mean rating for a given Work Style dimension (across the five experts).

Of prime interest in Table 3.2 are the ICC(C,5) values, as these reflect the reliability of the *mean expert ratings* that serve as the actual gold standard ratings for each Work Style dimension. Focusing on these values, we see that ICC(C,5) values range from .57 (Dependability) to .93 (Empathy, Intellectual Curiosity) – average of .84 across all Work Styles. For all but Dependability, these values indicate acceptable levels of interrater reliability for the mean expert ratings. For Dependability, the low reliability estimate reflects a lack of true score variance in Dependability across occupations (i.e., a real restriction in the range of ratings across occupations), as experts rated Dependability as "beneficial" or "very beneficial" for 95% of the sampled occupations. As such, any inconsistencies in expert ratings of Dependability across occupations would have an outsized impact on interrater reliability, given the low true score variance, even if those errors are very small in absolute terms.[7]

[7] Note, we confirmed this by examining variance components underlying the calculation of the ICCs in Table 3.2. Specifically, we found the occupation-variance component (i.e., the component corresponding to true score variance in the ICC calculations) for Dependability was far smaller than it was for other Work Styles, yet the residual variance component (i.e., a component that contributes to error variance in the ICC calculations) for Dependability was comparable to or smaller than it was for other Work Styles.

Table 3.2. Interrater Reliability and Agreement for Work Style Dimensions

Work Style	ICC(C,1)	ICC(C,5)	ICC(A,1)	ICC(A,5)
Empathy	.72	.93	.68	.91
Intellectual Curiosity	.72	.93	.61	.89
Optimism	.70	.92	.62	.89
Social Orientation	.69	.92	.65	.90
Leadership Orientation	.68	.91	.66	.91
Innovation	.65	.90	.61	.88
Cooperation	.62	.89	.50	.83
Self-Control	.57	.87	.53	.85
Cautiousness	.55	.86	.49	.83
Achievement Orientation	.54	.85	.43	.79
Integrity	.54	.85	.40	.77
Sincerity	.53	.85	.40	.77
Self-Confidence	.52	.85	.34	.72
Tolerance For Ambiguity	.52	.84	.52	.84
Adaptability	.50	.84	.45	.80
Stress Tolerance	.48	.82	.43	.79
Attention To Detail	.44	.80	.35	.73
Humility	.40	.77	.36	.74
Initiative	.38	.76	.29	.67
Perseverance	.38	.75	.34	.72
Dependability	.21	.57	.12	.42

Note. *N* = 125 (unit of analysis is occupation). Occupations are treated as targets of measurement. Work Styles are sorted in descending order of ICC(C,5) reliability estimates.

Beyond the reliability of ratings for the individual Work Style dimensions, we also examined the reliability (i.e., consistency) and absolute agreement (i.e., interchangeability) of the within-occupation Work Style profiles for expert ratings. 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 125 occupations where Work Style dimensions serve as the target of measurement, that is, we address how consistent expert raters are in terms of their ordering (and absolute rating) of Work Style dimensions for any given occupation. Table 3.3 shows the results of our analysis. Specifically, it provides descriptive statistics summarizing ICC statistics calculated for each of the 125 occupations. As with the findings above, the mean Work Style profiles provided by experts exhibit good levels of reliability (Mean ICC[C,5] = .86), with a 5th percentile of the ICC(C,5) value of .74.

Table 3.3. Reliability and Agreement for Within-Occupation Work Style Profiles

Statistic	ICC(C,1)	ICC(C,5)	ICC(A,1)	ICC(A,5)
Mean	.57	.86	.51	.83
SD	.12	.06	.13	.08
Min	.24	.62	.17	.51
Max	.74	.94	.73	.93
Percentiles				
5	.36	.74	.28	.66
25	.49	.83	.43	.79
50	.60	.88	.53	.85
75	.66	.91	.61	.89
95	.73	.93	.68	.91

Note. Separate ICC estimates were calculated for each occupation, where the *n* for each estimate was 21 (i.e., the number of Work Styles for an occupation). The reported statistics are summaries of ICC estimates across 125 occupations.

Correlations Among Work Styles

Lastly, we examined correlations among Work Styles to evaluate whether they were empirically distinct based on expert ratings (i.e., they did not exhibit perfect or near perfect correlations). A review of Work Style correlations in Table 3.4 reveals that no pair of Work Styles exhibited correlations of .90 or greater in magnitude, which would be suggestive of empirical redundancy among the pair considered. Three of 210 pairs of Work Styles (1.4%) did have correlations of .85 or above in magnitude, those being:

- Self-Control and Stress Tolerance (r = .89)
- Empathy and Optimism (r = .88)
- Optimism and Social Orientation (r = .87)

Out of all 210 pairs of Work Styles, 189 pairs (90.0%) had correlations of less than .70 in magnitude, and the average correlation was .35 (*SD* = .29, *Min* = -37 [Cautiousness, Social Orientation, *Max* = .89 [Self-Control, Stress Tolerance]).

As a follow-up to the analyses above, we partitioned variance in the set of potentially redundant Work Style pairs above to estimate the expected percentage of covariance that simply reflected the fact that the same set of raters provided both Work Style ratings. The observed correlation between Work Styles, as reported in Table 3.4, is not only a function of the "true score" correlation among Work Styles, but also the fact that Work Styles covary simply because of shared rater-specific error. These analyses revealed that 7.7% of the covariance among Self-Control and Stress Tolerance was attributable to those dimensions sharing a common set of raters, with 2.8% of the covariance among Empathy and Optimism attributable to a common set of raters, and 1.4% of the covariance among Optimism and Social Orientation.

Table 3.4. Work Style Intercorrelations based on Expert Ratings

Work Style		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Achievement Orientation	1.00																				
2	Adaptability	**.66**	1.00																			
3	Attention To Detail	.17	**.21**	1.00																		
4	Cautiousness	**-.23**	-.06	**.44**	1.00																	
5	Cooperation	.11	**.39**	-.14	-.11	1.00																
6	Dependability	-.14	.10	.12	**.41**	**.40**	1.00															
7	Empathy	.09	**.30**	**-.23**	-.13	**.84**	**.30**	1.00														
8	Humility	**-.23**	-.01	.00	**.19**	**.49**	**.36**	**.51**	1.00													
9	Initiative	**.54**	**.65**	-.12	**-.26**	**.54**	.08	**.43**	.16	1.00												
10	Innovation	**.66**	**.73**	.17	**-.28**	.15	**-.29**	.11	-.03	**.60**	1.00											
11	Integrity	**.33**	**.44**	**.27**	**.24**	**.44**	**.42**	**.39**	**.29**	**.37**	**.22**	1.00										
12	Intellectual Curiosity	**.64**	**.67**	**.46**	.03	.10	-.12	.09	.01	**.43**	**.80**	**.37**	1.00									
13	Leadership Orientation	**.57**	**.60**	-.02	-.06	**.52**	**.22**	**.46**	-.01	**.57**	**.36**	**.55**	**.39**	1.00								
14	Optimism	**.19**	**.35**	**-.31**	**-.20**	**.80**	**.27**	**.88**	**.42**	**.55**	.14	**.39**	.06	**.54**	1.00							
15	Perseverance	**.62**	**.65**	.02	-.16	**.36**	.07	**.36**	-.04	**.58**	**.47**	**.44**	**.43**	**.62**	**.42**	1.00						
16	Self-Confidence	**.64**	**.67**	.11	-.14	**.38**	.05	**.37**	-.16	**.51**	**.48**	**.49**	**.49**	**.74**	**.47**	**.70**	1.00					
17	Self-Control	**.18**	**.48**	-.04	.07	**.75**	**.44**	**.74**	**.29**	**.49**	.12	**.49**	.16	**.64**	**.77**	**.53**	**.60**	1.00				
18	Sincerity	**.20**	**.34**	-.12	.02	**.79**	**.36**	**.84**	**.60**	**.47**	.17	**.52**	**.21**	**.52**	**.82**	**.32**	**.42**	**.70**	1.00			
19	Social Orientation	**.38**	**.48**	**-.28**	**-.37**	**.81**	.12	**.78**	**.26**	**.63**	**.30**	**.41**	.16	**.66**	**.87**	**.50**	**.61**	**.72**	**.73**	1.00		
20	Stress Tolerance	**.28**	**.56**	.10	.16	**.69**	**.44**	**.63**	**.20**	**.50**	**.18**	**.51**	**.25**	**.63**	**.63**	**.58**	**.63**	**.89**	**.57**	**.60**	1.00	
21	Tolerance For Ambiguity	**.60**	**.77**	.13	-.14	**.42**	.00	**.37**	.06	**.67**	**.68**	**.49**	**.64**	**.59**	**.40**	**.72**	**.69**	**.50**	**.41**	**.51**	**.55**	1.00

Note. N = 125 (unit of analysis is occupation). Correlations that are bolded are statically significant (p < .05, two-tailed).

Step 4: Exploration of Alternative Rating Methods

Though the focus of our effort was on developing an efficient, accurate generative AI-based method for producing Work Style ratings, as part of Step 4, we initially identified and considered a wide variety of potential alternative methods for generating Work Style ratings that would not require future data collections with either job incumbents, occupational experts (OEs) or occupational analysts to maintain. That is, we explored methods that could use existing O*NET data from other parts of the O*NET Content Model to draw inferences about the impact of Work Styles on job performance in an occupation. The reason for considering multiple alternative methods was that, prior to this effort, it was unclear whether a generative AI-based approach would sufficiently mirror expert ratings. Therefore, there was interest in exploring alternative options in case the generative AI methods proved insufficiently performant. In the sections that follow, we provide a high-level summary of the methods we explored, the criteria on which different methods were evaluated, and conclude with a more detailed description of the best-bet method that emerged from this process, namely a composite of Work Style ratings generated by three different Large Language Models (LLMs).

Overview of Rating Methods Considered

We developed and evaluated three sets of alternative rating methods. The first set of methods was based on Work Style-Generalized Work Activity (WS-GWA) and Work Style-Work Context (WS-WC) linkage data collected during "Phase 1" of the O*NET Work Styles taxonomy revision effort (see Putka et al., 2024). The second method involved using SME judgment to identify other elements of the O*NET Concept Model that are most conceptually related to Work Styles, and using importance ratings of those elements as input to a regression model designed to predict expert ratings for each Work Style. The final set of methods was all LLM-based, where we explored different prompting strategies to identify the best option for generating ratings that mirrored those of experts. Each method was assessed against various evaluation criteria, and the highest-performing method was selected as the best-bet rating method that was further evaluated in subsequent steps of this effort.

Development of Linkage-Based Rating Methods

The first set of methods we developed leverages SME-based linkages between Work Styles and Generalized Work Activities (GWA) and Work Contexts from the first phase of the Work Styles revision effort. The logic behind these methods is as follows: For any given data-level occupation in O*NET, important GWAs and Work Contexts can be identified. For each of those important GWAs and Work Contexts, the linkage data described above indicates which Work Styles are expected to impact (positively or negatively) the performance of those activities and performance in those contexts. Given such information, one set of methods for drawing inferences about the impact a given Work Style may have on the performance of an occupation in general would be to infer it based on whether the given Work Style impacts GWAs and Work Contexts that are important to that occupation. Based on this concept, we formulated nine different ways we could use the WS-GWA and WS-WC linkage data described above to estimate the impact a Work Style may have on performance in an occupation. The methods we developed are listed in Table 4.1. All ratings generated via these methods were based on data from O*NET 29.0 (i.e., the same version of O*NET underlying the expert and analyst data collections), with ratings assigned to each Work Style for all 125 occupations rated by experts and analysts.

Table 4.1. Description of Linkage-Based Rating Methods

Rating Method	Rating Method Label	Description
RM1	Simple Average	Mean WS-WC linkage rating for the top seven rated Work Contexts.
RM2	Percentage Ratio	Percentage of positive WS-WC linkages for the top seven rated Work Contexts.
RM3	TF-IDF Weighted Average	Mean of WS-WC linkage ratings for the top seven TF-IDF weighted important Work Contexts.
RM4	TF-IDF Weighted Percentage Ratio	Percentage of positive WS-WC linkages for the top seven TF-IDF weighted important Work Contexts.
RM5	Simple Average	Mean WS-GWA linkage rating for the top seven rated GWAs.
RM6	Percentage Ratio	Percentage of positive WS-GWA linkages for the top seven rated GWAs.
RM7	TF-IDF Weighted Average	Mean WS-GWA linkage rating for the top seven TF-IDF weighted important GWAs.
RM8	TF-IDF Weighted Percentage Ratio	Percentage of positive WS-GWA linkages for the top seven TF-IDF weighted important GWAs.
RM9	Combined Average	Mean Importance of GWA-WC with strong positive links to WS.

Rating Methods (RMs) 1 and 5 computed Work Style ratings as simple averages of linkage ratings (i.e., impact of a given Work Style on performance on a given GWA or in a given Work Context) across the top seven most important GWAs (RM5) or Work Contexts (RM1) for an occupation. RMs 2 and 6 calculated the percentage of positive linkages between Work Styles and the top seven most important GWAs (RM6) or Work Contexts (RM2) as a rating for each occupation.

One drawback of the methods above is that GWAs and WCs were rated as important across most occupations, making simple averages of top-rated elements potentially indistinguishable across different occupations. To address this limitation, we implemented a term frequency-inverse document frequency (TF-IDF) weighting scheme to prioritize GWAs and Work Contexts that were identified as important to the target occupation, but that were identified as important less frequently across occupations. In our case, term frequency (TF) represented an indicator of whether the given GWA/Work Context was important to the target occupation (1 = Yes, important; 0 = No, not important), while Inverse Document Frequency (IDF) reflected the proportion of occupations in the database where the given GWA/Work Context was identified as important. We applied TF-IDF weighting to the importance ratings of GWAs/Work Contexts for each occupation and re-selected the top 7 most important GWAs/Work Contexts for an occupation after weighting. RMs 3, 4, 7, and 8 represent the TF-IDF weighted variants of RMs 1, 2, 5, and 6, respectively. Lastly, RM 9 integrated both WS-GWA and WS-WC linkages by calculating the mean importance ratings of GWAs and Work Contexts that demonstrated strong positive links (mean linkage ratings of at least 2.0 on the -3 to +3 scale) to each Work Style.

Development of a Rational Regression-Based Composite Rating Method

Taking advantage of the diverse set of variables available in the O*NET Content Model, we also explored the use of multiple regression models to predict Work Style ratings. To identify elements of the O*NET Content Model most pertinent to predicting Work Styles, the HumRRO

team, including four PhD-level I-O Psychologists, conducted a rational exercise to identify elements of the O*NET Content Domains believed most likely to predict each Work Style.

The candidate set of O*NET elements included: RIASEC interest dimensions; six GWA-Work Context components derived from Principal Component Analysis (PCA) of GWA (importance) and Work Context ratings; four skill components from PCA of skill importance ratings; five ability components from PCA of ability importance ratings; and selected individual Work Contexts, skill, and GWA elements that demonstrated direct theoretical correspondence with specific Work Styles.

The exercise began with the independent nomination of the most relevant elements by each participating HumRRO staff member for each Work Style separately, followed by consensus-building discussions to determine what were believed to be the best three elements to include in a prediction model (considering also the ability of the set of elements to complement each other within the context of the model). This process resulted in the identification of the top three GWA-WC-Ability-Skill-Interest elements judged most relevant for use in a prediction model for each Work Style. These elements provided the basis for Rating Method 10, which reflects a regression-based composite of the top three elements above for each Work Style, where each element's weight was estimated using an Ordinary Least Squares (OLS) regression of the final expert ratings for a given Work Style (from Step 3) on these three rationally selected elements. Separate models were fitted for each of the 21 Work Styles on the full set of 125 occupations for which experts and analysts provided ratings.

Development of LLM-Based Rating Methods

Our final and focal set of methods involved prompting LLMs to generate Work Style ratings. These methods included prompts that integrated Work Style descriptors (e.g., name and definition) with occupation-specific information (e.g., occupation name, description, key tasks, skills, and relevant GWAs and Work Contexts). One set of LLM methods we explored employed a zero-shot prompting strategy, presenting information without reference points or rating examples. The second set of LLM methods we explored implemented a few-shot prompting strategy that incorporated examples and anchors for different rating levels, such as examples of types of work activities that reflect high, moderate, or low standing on the given Work Style, and examples of occupations that fall at each rating point for a given Work Style.

Input Data Preparation

To prepare input data for LLM prompts, we developed a dataset containing occupational information (drawn from O*NET 29.0) for the subset of 125 occupations rated by experts and analysts. The set of information for each occupation included: job title, description, tasks, skills, Work Contexts, and job zone description, along with names and descriptions of each Work Style.

The tasks for each occupation were limited to core tasks, except for those occupations where core tasks could not be identified (due to a lack of task ratings for the occupation in O*NET 29.0); for these latter occupations, all associated tasks were included. The top ten skills for each occupation (based on importance ratings) were included. As noted earlier, given that many Work Contexts received very high importance ratings, we only included the most distinctive yet important Work Contexts for an occupation (based on TD-IDF rankings described in Step 3). For each occupation, tasks, skills, and work contexts were ordered by descending order of their importance ratings.

Platform for Prompt Development

To facilitate the prompt development process and automate LLM-based rating generation at scale, we employed HumRRO's secure generative AI platform (RRObot), which is built upon Amazon Web Services (AWS) Bedrock service (https://aws.amazon.com/bedrock/). AWS Bedrock is a service hosted by AWS that provides organizations with access to a variety of foundation models (i.e., LLMs from third-party providers such as Anthropic, Meta, OpenAI, and Amazon) to drive the functionality of generative AI applications, including HumRRO's RRObot platform (Human Resources Research Organization, 2025). The RRObot platform provides HumRRO staff with a suite of AI tools designed to facilitate HumRRO's provision of high-stakes assessment and talent management solutions. RRObot operates in a completely secure environment (based on our current National Institute of Standards and Technology [NIST]/International Organization for Standardization [ISO]/Cybersecurity Maturity Model Certification [CMMC] compliance and boundaries) where client information remains protected. RRObot allows HumRRO staff to interact with LLMs. It provides HumRRO staff with highly accessible and practical means to significantly streamline, enhance, and re-imagine the way we do our work, leveraging generative AI.

For this effort, we leveraged RRObot's batch processing capabilities to upload a file of (a) prompt templates designed to produce ratings for a given Work Style-occupation combination that took on information on the specific Work Style and occupation to be rated as variable inputs, (b) information on the occupations to be rated (e.g., job descriptions, task statements, work contexts), (c) information on the Work Styles to be rated (e.g., names and descriptions of the Work Styles), and (d) configuration information that identifies LLMs and hyperparameters to be used when generating ratings. The application automatically populates the templates with information for each Work Style-occupation combination, creating populated prompts for every combination. These prompts are then batch-processed through LLMs to generate ratings for each combination.

Large Language Models and Hyperparameters

For this effort, we initially focused our experimentation on a set of lightweight, fast, and cost-effective LLMs, along with one model representing the most advanced option available within the RRObot environment at the time. Specifically, the initial models we experimented with included Claude 3 Haiku, Llama 3 8B, Mistral 7B, Mixtral 8x7B, and Claude 3.5 Sonnet v1. Table 4.2 summarizes the LLMs examined and used throughout the project.

We first developed and refined prompt templates within RRObot's chat environment, enabling real-time interaction with the first five LLMs listed in Table 4.2. This setup ensured that prompts functioned as intended (e.g., producing ratings within the valid range and in the specified format). We then experimented with key hyperparameters to improve the prompts' functioning. Specifically, we varied: (1) temperature (controlling randomness in token selection) and (2) top p (nucleus sampling parameter, typically used with temperature). After preliminary testing, we standardized hyperparameters across all prompts and models by setting temperature = 0.2 and top p = 0.2.

We next compared model performance with respect to how well they produced different ratings for contrasting occupations. At the beginning of the experiment stage, expert data collection was still ongoing; therefore, expert and analyst ratings were not yet available. As a proxy, we utilized published O*NET Work Styles ratings. We selected a sample of 50 occupations across four Work Styles (Social Orientation, Cooperation, Humility, and Innovation) that were previously

included in the O*NET Work Styles taxonomy developed by Borman and colleagues. Our application automatically generated and executed prompts for all 200 Work Style-occupation combinations (50 occupations × 4 Work Styles) for each of the LLMs with which we experimented.

Table 4.2. List of LLMs Examined and Used

LLM	Full LLM Name-Version	Release Date	# of Parameters	Usage	Citation
Claude 3 Haiku	anthropic.claude-3-haiku-20240307-v1:0	2024-03-13	Not disclosed	Used for experimentation	Anthropic (2024)
Llama 3 8B	meta.llama3-8b-instruct-v1:0	2024-04-18	8B	Used for experimentation	Duby et al. (2024)
Mistral 7B	mistral.mistral-7b-instruct-v0:2	2023-09-27	7B	Used for experimentation	Jiang et al. (2023)
Mixtral 8x7B	mistral.mixtral-8x7b-instruct-v0:1	2023-12-11	56B	Used for experimentation	Jiang et al. (2024)
Claude 3.5 Sonnet v1	anthropic.claude-3-5-sonnet-20240620-v1:0	2024-06-20	Not disclosed	Used for experimentation and final rating generation	Anthropic (2024)
Claude 3.5 Sonnet v2	anthropic.claude-3-5-sonnet-20241022-v2:0	2024-10-22	Not disclosed	Used for final rating generation	Anthropic (2024)
Llama 3.3 70B Instruct	llama-3.3-70b-instruct-maas	2024-12-06	70B	Used for final rating generation	Duby et al. (2024)

During evaluation, the two Mistral-Mixtral models (Mistral 7B and Mixtral 8x7B) produced a substantial number of missing values due to output formatting issues that prevented the successful automated extraction of ratings. As a result, these two models were excluded from further consideration. Among the remaining models, Claude 3 Haiku and Llama 3 8B exhibited comparable performance, with moderate convergence to the published Work Styles ratings. However, Claude 3.5 Sonnet v1 demonstrated superior performance in terms of both accuracy and consistency (with published ratings) across the evaluated dimensions. This initial exploration indicated that while smaller models such as Claude 3 Haiku and Llama 3 8B were suitable for quick, lower-cost tasks, Claude3.5 Sonnet v1 provided more reliable performance for our rating task. Based on these results, among the options we initially experimented with, we narrowed our focus to Claude 3.5 Sonnet v1 for subsequent evaluation and rating generation. As we describe below, considering advances in LLMs over the course of this project, we wound up considering additional LLMs for further evaluation as we strived to take advantage of new, potentially better models for the rating task at hand as they became available. Specifically, Claude 3.5 Sonnet v2 and Llama 3.3 70B Instruct (a larger version of Llama 3.8 B) were added to our LLM list when they became available for evaluation.

Prompt Development

Prior to evaluating different prompting strategies using the full set of 125 occupations rated by experts and analysts, we evaluated the impact of the differential placement of job input within the prompt, as well as the rating scale, on prompt performance. For job input placement, we tested two placement positions: (1) positioning job information at the beginning of the prompt

immediately following the introductory statement, and (2) positioning job information at the end of the prompt after the rating scale description. We tested two rating scales: (1) a standard 1–7-point scale (where 1 = very detrimental and 7 = very beneficial) and (2) an expanded 5-35 point scale (linearly mapping the original 1-7 scale to 5 = very detrimental and 35 = very beneficial). The rationale behind using the 5-35 scale was to address a scale coarseness issue by asking LLMs for ratings on a 7-point rating scale. Specifically, with a 7-point scale, LLMs tend to limit returned ratings to whole numbers (e.g., 1, 2, 3, 4, 5, 6, or 7). In contrast, the expert ratings we have available for evaluation offer a far more fine-grained scale, as when 1-7 ratings are averaged across five experts, intermediate values can result (e.g., 1.2, 1.4, 1.6). We wanted to ensure that the ratings generated by LLM could were as fine-grained as the average expert ratings, so our approach to doing this was simply to increase the number of scale points such that we can easily divide the resulting 5-35 point ratings by five to arrive back at a set of ratings on a 1-7 scale that mirrors the level of specificity in average expert ratings.

To assess these design variations, we conducted another round of pilot testing using the same sample of 50 occupations identified in the previous step across four Work Styles (Social Orientation, Cooperation, Humility, and Innovation). Using Claude 3.5 Sonnet v1, our application automatically generated and executed prompts for all 200 Work Style-occupation combinations. We evaluated prompt performance against expert benchmarks using two metrics: (1) absolute agreement between LLM-generated ratings and expert mean ratings for each Work Style, and (2) correlation between LLM and expert ratings for each Work Style. Based on this evaluation, we identified a prompt design that strikes a balance between effectiveness and parsimony during the early testing stage. The selected baseline prompt used a 5-35 point rating scale and positioned job information at the beginning of the prompt, serving as the baseline template for subsequent development and refinement.

Building on the baseline prompt, we developed and refined eight distinct prompt variations across zero-shot and few-shot prompting methods. A detailed description and comparison of all eight LLM prompts are presented in Table 4.3.

First, we developed and refined four zero-shot prompts that varied in their job input elements, incorporating different combinations of job tasks, skills, Work Contexts, and job zone information. Second, we developed four few-shot prompts using the same baseline template structure. The few-shot variations differed primarily in whether they provided exemplar occupations and/or GWA/Work Contexts associated with high, middling, or low ratings on the Work Style rating scale.

To identify few-shot occupation examples, we selected two to three occupations for each Work Style based on expert ratings: occupations where the Work Style was rated as having a beneficial impact (high set), having little or no impact (middling set), or having a detrimental impact (low set) to job performance, respectively. For Work Styles that were consistently rated as beneficial across most occupations, no or few exemplar occupations were provided as low or middling few-shot examples due to the insufficient number of examples in those ranges.

We applied a similar approach to identify few-shot GWA/Work Context examples. Specifically, utilizing the WS-GWA and WS-WC linkage data collected during Phase 1 of the Work Styles revision effort, for each Work Style, we selected two to three GWAs or Work Contexts that were rated as having a beneficial impact (high set), having little or no impact (middling set), and having a detrimental impact (low set). For Work Styles that were consistently rated as beneficial across most GWAs or Work Contexts, no or few exemplar GWAs or Work Contexts were provided for as low or middling few-shot examples due to insufficient examples in those ranges.

Table 4.3. Description of LLM-Based Methods

Rating Method	Prompt Type	Basic Job Input	Additional Job Input	Few-shot Example
RM11	Zero-shot	Name, description, and tasks	None	None
RM12	Zero-shot	Name, description, and tasks	Top 10 most important skills	None
RM13	Zero-shot	Name, description, and tasks	Top 10 most important skills, and distinguishing Work Contexts	None
RM14	Zero-shot	Name, description, and tasks	Top 10 most important skills, distinguishing Work Contexts, and job zones	None
RM15	Few-shot	Name, description, and tasks	None	Occupation examples at low, mid, and high levels
RM16	Few-shot	Name, description, and tasks	Top 10 most important skills	Occupation examples at low, mid, and high levels
RM17	Few-shot	Name, description, and tasks	None	Occupation and GWA/WC examples at low, mid, and high levels
RM18	Few-shot	Name, description, and important tasks	Top 10 most important skills	Occupation and GWA/WC examples at low, mid, and high levels

With the prompt template, job input, Work Style input, and configuration file constructed, we used the RRObot batch to automatically populate and run prompts for 21 Work Style x 125 occupation combinations through Claude Sonnet 3.5 v1, the most performant model within our system at the time of research. These runs yielded a set of eight ratings for each Work Style dimension for each occupation (i.e., one rating each for RM 11 through RM 18).

Initial Evaluation of Alternative Methods

We evaluated the extent to which ratings from each alternative method aligned with expert ratings from Step 3 by estimating correlations between each method's ratings and expert ratings (treating occupations as units of analysis). Results are summarized in Table 4.4.

Among the 18 rating methods, the linkage-based methods (RMs 1–9) consistently exhibited lower correlations with expert ratings across Work Styles (average *r*s ranged from .38 to .56), compared to RM 10 (rational-regression-based method) and the LLM-based methods. In contrast, RM 10, the rational regression-based method, demonstrated comparable correlations (average $r = .71$) with expert ratings to the LLM methods using zero-shot prompts (average *r*s ranged from .69 to .70). The few-shot prompt methods (RMs 15-18) emerged as the best-performing approaches among all 18 empirical methods, with average correlations with expert ratings being .77 for all four prompts.

Given that the four few-shot prompts produced very similar performance in terms of rating distributions and correlations with expert ratings, it appeared that inclusion of additional job input elements and work activity examples in RMs 16-18 did not provide meaningful improvements over the prompt using basic job input and providing solely occupation few-shot examples (RM 15). Therefore, we moved forward with RM 15 as the best-bet rating method at this stage of analysis.

Table 4.4. Convergent Correlations Between Empirical Ratings from RMs 1-18 and Expert Ratings by Work Styles

Work Style	RM1	RM2	RM3	RM4	RM5	RM6	RM7	RM8	RM9	RM10	RM11	RM12	RM13	RM14	RM15	RM16	RM17	RM18
Achievement Orientation	0.37	0.30	0.60	0.54	0.59	0.53	0.55	0.45	0.65	0.57	0.66	0.63	0.62	0.63	0.74	0.76	0.74	0.75
Adaptability	0.51	0.48	0.54	0.48	0.51	0.49	0.44	0.35	0.61	0.73	0.65	0.62	0.65	0.67	0.76	0.72	0.77	0.79
Attention To Detail	0.30	0.13	0.35	0.31	0.47	0.49	0.39	0.33	0.34	0.71	0.57	0.66	0.66	0.66	0.77	0.72	0.74	0.68
Cautiousness	0.49	0.47	0.58	0.54	0.54	0.53	0.35	0.34	0.55	0.70	0.58	0.61	0.62	0.61	0.73	0.71	0.71	0.67
Cooperation	0.60	0.58	0.52	0.49	0.68	0.64	0.66	0.63	0.70	0.81	0.81	0.82	0.78	0.77	0.83	0.81	0.86	0.81
Dependability	0.00	-0.05	-0.06	-0.07	0.33	0.40	0.20	0.26	0.51	0.42	0.53	0.44	0.54	0.49	0.51	0.61	0.58	0.51
Empathy	0.51	0.54	0.51	0.55	0.69	0.60	0.69	0.65	0.73	0.88	0.85	0.85	0.83	0.84	0.86	0.88	0.85	0.85
Humility	0.23	0.21	0.15	-0.11	0.20	0.20	0.14	0.14	NA	0.47	0.50	0.48	0.51	0.46	0.63	0.66	0.67	0.69
Initiative	0.55	0.54	0.50	0.42	0.46	0.42	0.47	0.25	0.39	0.52	0.60	0.59	0.62	0.59	0.69	0.70	0.69	0.68
Innovation	0.50	0.53	0.65	0.60	0.70	0.56	0.62	0.50	0.70	0.81	0.84	0.83	0.83	0.82	0.86	0.87	0.89	0.87
Integrity	0.50	0.41	0.49	0.47	0.37	0.19	0.52	0.39	0.29	0.64	0.67	0.66	0.68	0.66	0.75	0.73	0.70	0.78
Intellectual Curiosity	0.49	0.41	0.60	0.49	0.68	0.68	0.55	0.58	0.63	0.88	0.86	0.84	0.85	0.86	0.88	0.85	0.86	0.87
Leadership Orientation	0.68	0.59	0.66	0.57	0.63	0.66	0.64	0.62	0.62	0.85	0.84	0.86	0.85	0.87	0.86	0.87	0.88	0.88
Optimism	0.55	0.53	0.55	0.48	0.71	0.65	0.64	0.62	0.64	0.84	0.81	0.79	0.76	0.76	0.80	0.83	0.80	0.79
Perseverance	-0.30	-0.34	-0.04	-0.02	0.11	0.04	0.01	-0.13	0.33	0.55	0.52	0.50	0.53	0.57	0.63	0.66	0.63	0.62
Self Confidence	0.61	0.57	0.62	0.59	0.57	0.48	0.59	0.57	0.66	0.81	0.72	0.67	0.66	0.71	0.80	0.78	0.80	0.80
Self Control	0.45	0.44	0.52	0.43	0.50	0.55	0.47	0.52	0.68	0.74	0.75	0.72	0.74	0.67	0.76	0.75	0.78	0.81
Sincerity	0.49	0.46	0.52	0.42	0.57	0.54	0.62	0.62	0.70	0.81	0.69	0.66	0.65	0.64	0.72	0.70	0.71	0.76
Social Orientation	0.66	0.66	0.65	0.68	0.72	0.69	0.73	0.71	0.74	0.82	0.86	0.88	0.89	0.86	0.91	0.92	0.91	0.89
Stress Tolerance	0.08	0.10	0.39	0.39	0.32	0.52	0.33	0.47	0.28	0.69	0.71	0.74	0.66	0.59	0.75	0.78	0.79	0.77
Tolerance For Ambiguity	0.46	0.41	0.55	0.43	0.52	0.46	0.49	0.46	0.45	0.62	0.68	0.65	0.74	0.72	0.83	0.80	0.80	0.82
Average	0.41	0.38	0.47	0.41	0.52	0.49	0.48	0.44	0.56	0.71	0.70	0.69	0.70	0.69	0.77	0.77	0.77	0.77

Note: n = 125 (unit of analysis is occupation). RM 9 was based on the average importance ratings of GWA-WC linkages that had a mean linkage rating of at least 2.0 (on the -3 to 3 scale) with a given Work Style. All Work Styles had at least one GWA–WC linkage meeting this threshold, except for Humility. As a result, the convergent correlation for Humility in RM 9 could not be calculated (NA).

Refinement of Best-Bet LLM Rating Method

To refine our working best-bet rating methods (RM 15), we conducted further experimentation by examining four additional variations of the RM 15 prompt (see Table 4.5).

Table 4.5. Description of LLM-Based Methods

Rating Method	Description
RM15	Baseline
RM19	Same prompt as ERM 15, but using a rating scale from 5-25.
RM20	Based on ERM 15, but using a rating scale from 5-25, and incorporated a chain-of-thought prompting strategy.
RM21	Variation of ERM 20 using “occupation” as the primary reference, and added examples of “job titles” for a given occupation as an additional job input.
RM22	Based on ERM 15, but providing seven anchors on a rating scale from 5-35, and used custom few-shot occupation examples.

The first revision involved collapsing the rating scale from 7 points to 5 points. Specifically, ratings of -3, -2, and -1 on the impact rating scale metrics were converted to 1 (*detrimental to performance*), while non-negative ratings of 0 to +3 were linearly mapped to 2 (*has little or no impact*) through 5 (*very beneficial to performance*). This collapsing was done due to the relative scarcity of negative ratings. We then updated the scaling in the prompt to reflect this change. Specifically, RM 19 used the same prompt structure as RM 15 but employed a 5-25 point rating scale. RM 20 retained the 5-25 point scale and incorporated a chain-of-thought prompting strategy. The idea of RM 20 was to prompt LLMs to break down complex questions into smaller and intermediate steps, which might help in arriving at more accurate results (Wei et al., 2022). RM 21 further modified RM 15 by using "occupation" as the primary referent rather than "job" and introducing example job titles for each occupation as additional input.

The second major revision modified the few-shot examples used in the prompt. Rather than providing occupation exemplars only at low, middling, and high impact rating levels (where available), we sampled occupations at multiple anchor points across the 5–35 rating scale, using expert ratings from Step 3 for each Work Style. Specifically, we first transformed expert mean ratings from the original −3 to +3 scale to the equivalent 5–35 scale by first adding 4 to the −3 to +3 scale and then multiplying the scale points by 5. We then partitioned each Work Style’s rating distribution into five to seven data ranges. As noted in Step 3, most Work Styles generally had “positive” impact on occupational performance (e.g., expert means > 0 on the original scale, i.e., > 20 on the 5–35 scale). Consequently, not all Work Styles had anchors covering the full 5–35 range; where the empirical distribution lacked coverage (e.g., at the low end), those anchor positions were intentionally left unfilled rather than fabricated. After defining the data ranges for each Work Style, we sampled three to five occupations from each populated range. Because occupations that were representative of different rating levels varied by Work Style, each Work Style received its own set of few-shot anchors. This process yielded our final prompt variant, RM 22 (based on RM 15), which provides Work-Style-specific few-shot occupation examples on the 5–35 scale while reflecting the actual distributional coverage available for each Work Style.

To minimize rating idiosyncrasies inherent in the stochastic nature of LLMs, we ran baseline RM 15 and the four variations (RMs 19-22) three times using Claude Sonnet 3.5 v1. Final ratings for each prompt represented average scores across runs. We then compared the four variations against RM 15 using the same evaluation criteria as in the prior steps (i.e., correlations with

expert ratings, see Table 4.6). Results revealed that running prompts multiple times and averaging ratings across runs yielded higher correlations with expert ratings. However, prompt variations reflected in RMs 19-21 offered no advantage over RM 15. If anything, truncating to a 5-25 point scale actually slightly diminished prediction accuracy. Conversely, providing tailored few-shot examples in RM 22 yielded superior predictions compared to the baseline in RM 15. Note that some of this advantage may stem from the fact that the expert ratings themselves informed few-shot occupation examples used in each Work Style's RM 22 prompt, but at this point, the goal was to establish the best-prompt possible for generating Work Style ratings as much as possible, to include leveraging information from expert ratings.[8]

Table 4.6. Convergent Correlations Between Empirical Ratings from RMs 15, 19-22 and Expert Ratings by Work Styles

Work Style	RM15		RM19	RM20	RM21	RM22
Achievement Orientation	.74		.73	.73	.73	.83
Adaptability	.76		.76	.78	.75	.76
Attention To Detail	.77		.71	.64	.63	.87
Cautiousness	.73		.71	.66	.68	.81
Cooperation	.83		.79	.79	.76	.87
Dependability	.51		.55	.52	.57	.66
Empathy	.86		.85	.86	.87	.90
Humility	.63		.67	.65	.59	.70
Initiative	.69		.69	.68	.66	.72
Innovation	.86		.86	.88	.87	.92
Integrity	.75		.64	.65	.65	.80
Intellectual Curiosity	.88		.87	.84	.85	.90
Leadership Orientation	.86		.90	.90	.89	.94
Optimism	.80		.77	.83	.74	.85
Perseverance	.63		.63	.54	.62	.70
Self Confidence	.80		.79	.78	.75	.83
Self Control	.76		.68	.68	.62	.79
Sincerity	.72		.74	.71	.69	.78
Social Orientation	.91		.91	.91	.90	.94
Stress Tolerance	.75		.66	.69	.61	.81
Tolerance For Ambiguity	.83		.80	.81	.82	.88
Average	.77		.75	.74	.73	.82

Note: n = 125 (unit of analysis is occupation).

Based on the evaluation above, RM 22 emerged as the most effective method among all the tested approaches. The prompt template for ERM 22 is displayed in Appendix B along with

[8] In Step 5, we revisit this issue and provide an estimate of the cross-validity of RM 22-based ratings for predicting expert ratings that remove occupations used as few-shot examples from the sample of occupations used to calculate LLM-expert rating correlations. This is designed to address the critique that the results presented in Table 4.6 are overly inflated due to the use of expert ratings to inform the occupation examples used in the prompts underlying RM 22.

example inputs for each variable input field in the prompt template, and a table of the custom few-shot content for each Work Style.

In an effort to produce Work Style ratings produced by RM 22 even more robust, and less prone to both model-specific error (i.e., variation in ratings based on choice of LLM used) and the particular occasion on which the model is run (i.e., variation in ratings stemming from the fact LLMs are not deterministic and as such, generated ratings can slightly differ even with the same prompt and input), we took one final development step. Specifically, the RM 22 prompt template described in Appendix B was populated with information for each Work Style-occupation combination to be rated and subsequently run through three LLMs (Claude 3.5 Sonnet v1, Claude 3.5 Sonnet v2, and Llama 3.3 70B Instruct) on three separate occasions, yielding a total of nine impact ratings for the given Work Style-occupation combination. The decision to consider these models was driven by them being the latest models available within our generative AI platform at the time this work was conducted, and the general observation that newer general models tend to perform better than older generation models.[9] For all runs, the hyperparameter settings were .2 for temperature and .2 for top p. At this point, occupation inputs were still drawn from O*NET 29.0 (to maintain comparability with the information experts and analysts used when making their ratings). The nine ratings that resulted from the runs for each Work Style-occupation combination were rescaled to the -3 to +3 impact rating scale metric (to mirror the expert rating metric) by dividing the 5-35 scale by 5 and then subtracting 4. Next, the three re-scaled ratings for each model were averaged together to form a raw composite rating based on each LLM, and those composites were subsequently averaged together to form a raw Work Style rating for the occupation. As we will illustrate in our next evaluation step, by basing LLM-based Work Style ratings on multiple models and multiple runs, we were able to greatly reduce the amount of model-specific and run-specific error in Work Style ratings relative to if we had based the ratings on a single model, run on a single occasion.

[9] Note, we did confirm that adding in the additional two models improved correlations between RM 22-based ratings and expert ratings relative to if we had only based the LLM ratings on the LLM we had been using for evaluation to this point (i.e., Claude 3.5 Sonnet v1).

Step 5: Evaluation of LLM Ratings Against Expert Ratings

Our next step involved evaluating the best-bet LLM rating method emerging from Step 4 against expert ratings, which served as the gold standard. For comparison, we also evaluated analyst ratings alongside expert ratings. In the sections that follow, we summarize the results of the following evaluation-focused analyses:

- **Work Style rating distributions and descriptives by source**: We evaluated Work Style rating means, standard deviations, and distributions of expert, LLM, and analyst-based ratings for similarity/differences.
- **Reliability and agreement of Work Style ratings by source**: We compared Work Style rating reliability and absolute agreement estimates for expert, LLM, and analyst-based ratings.
- **Comparison of Work Style correlation matrices across sources:** We evaluated the extent of differences in Work Style rating intercorrelations among expert, LLM, and analyst-based ratings.
- **Convergence of Work Style ratings across sources:** We evaluated the level of correlation between expert and LLM ratings, and between expert and analyst ratings. Two types of correlations were examined. Between-occupation Work Style correlations that treated occupations as the units of analysis (i.e., separate correlations for each Work Style), and within-occupation Work Style profile correlations that treated Work Styles as the units of analysis (i.e., separate correlations for each occupation).
- **Within-occupation differences in Work Style ratings across sources**: We evaluated the extent of within-occupation differences between expert and LLM ratings and summarized those differences by job zone and job family to ascertain whether such differences varied as a function of those occupational groupings.

Work Style Rating Distributions and Descriptives by Source

We first examined smoothed frequency distributions of LLM, analyst, and expert ratings for each Work Style (see Figure 5.1), as well as the means and standard deviations of ratings from each rating source (see Table 5.1). Specifically, Figure 5.1 and Table 5.1 show distribution information for four sets of Work Style ratings:

- **Raw LLM Ratings (LLM-R)**: These are the LLM ratings that emerged from the best-bet method identified in Step 4 before any calibration with expert ratings was performed.
- **Calibrated LLM Ratings (LLM-C)**: These are the LLM ratings that emerged from the best-bet method identified in Step 4 following calibration against expert ratings. We created this variant based on observations that pre-calibrated LLM ratings tended to be higher and more negatively skewed than expert ratings, almost regardless of Work Style. This calibration involved two steps. First, we implemented a power transformation on the LLM ratings, designed to bring the skew of the more extreme ratings into line with the expert ratings for each Work Style. Second, we applied a mean/SD transformation to the resulting power-transformed LLM ratings, using the expert ratings as a norming sample. Appendix C provides further details on the final transformations applied to the pre-calibration LLM ratings to derive the calibrated LLM ratings, which effectively reflect hybrid generative AI-expert ratings.

- **Analyst Ratings**: These are the cleaned mean analyst ratings resulting from Step 2.
- **Expert Ratings**: These are the cleaned mean expert ratings resulting from Step 2.

As shown in Figure 5.1 and even more so in Table 5.1, calibrating the raw LLM ratings (LLM-R) had the intended effect of reducing differences between the resulting calibrated LLM ratings (LLM-C) and expert ratings, and confirming that the LLM-C and expert rating means and standard deviations for any given Work Style were the same post-calibration. Also of note in Figure 5.1 and Table 5.1, even prior to calibration, raw LLM ratings appeared to do a better job of differentiating among occupations than analyst ratings as evidenced by the observation that (a) standard deviations for the raw LLM ratings tended to be higher than the standard deviations for analyst ratings, and (b) raw LLM ratings tended to have flatter and less sharply peaked rating distributions relative to distributions for analyst ratings. Though the mean analyst ratings tended to be closer to the mean expert ratings (relative to the mean raw LLM ratings vs. the mean expert ratings), as noted above, the potential leniency bias in the raw LLM ratings can be remedied via the calibration with a representative calibration sample, which we had with the expert ratings.

Figure 5.1. Work Style Rating Distributions by Source

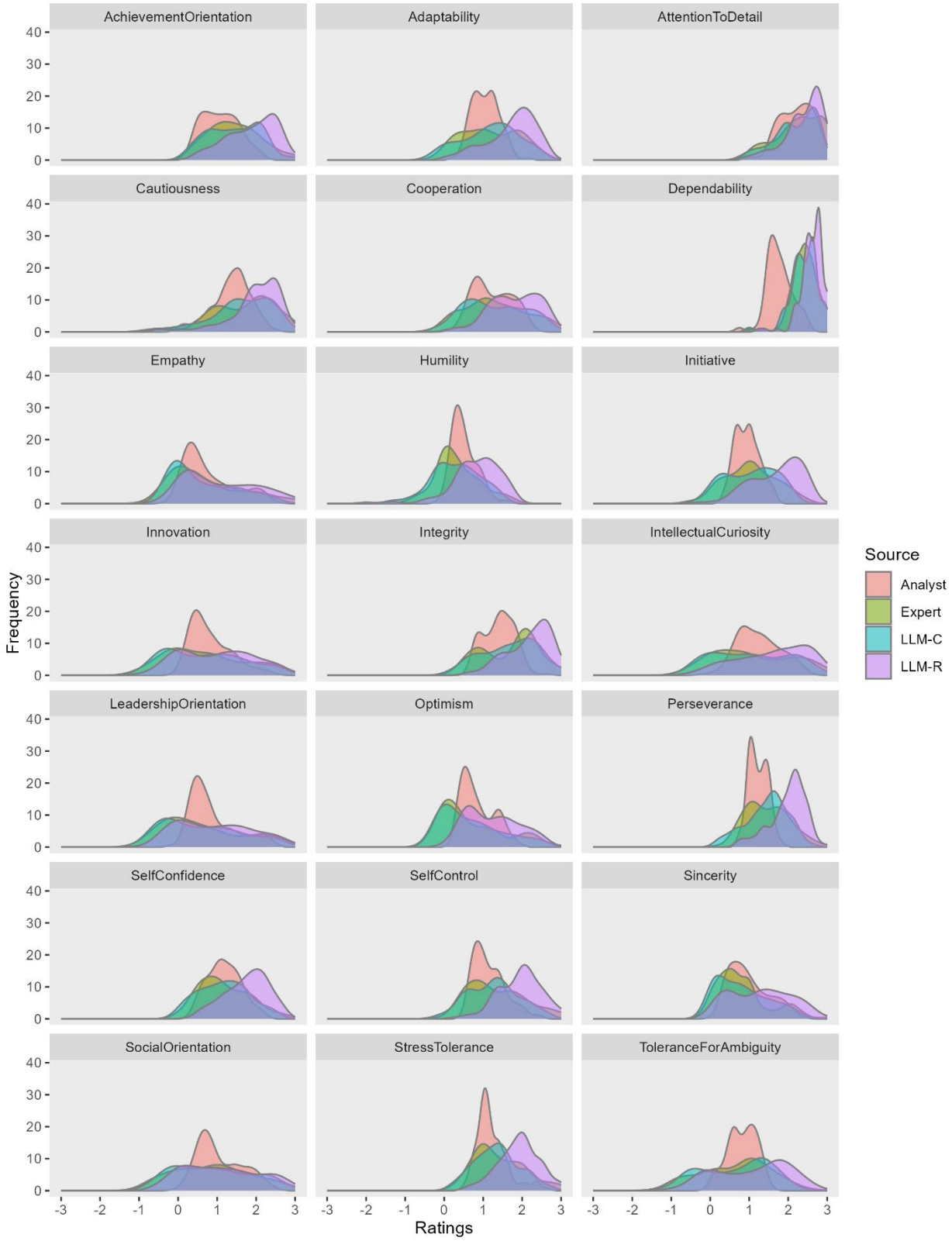


Table 5.1. Work Style Rating Descriptives by Source

Work Style	LLM-R		LLM-C		Analyst		Expert	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Dependability	2.61	0.29	2.41	0.35	1.77	0.32	2.41	0.35
Attention To Detail	2.41	0.48	2.27	0.58	2.21	0.42	2.27	0.58
Integrity	2.19	0.56	1.68	0.71	1.40	0.40	1.68	0.71
Cautiousness	2.01	0.66	1.67	0.82	1.37	0.50	1.67	0.82
Perseverance	2.00	0.46	1.47	0.56	1.23	0.24	1.47	0.56
Achievement Orientation	1.90	0.63	1.44	0.66	1.08	0.47	1.44	0.66
Stress Tolerance	1.92	0.52	1.41	0.64	1.22	0.40	1.41	0.64
Self-Control	1.95	0.55	1.34	0.72	1.15	0.41	1.34	0.72
Cooperation	1.82	0.62	1.32	0.79	1.24	0.49	1.32	0.79
Adaptability	1.75	0.62	1.19	0.72	1.06	0.34	1.19	0.72
Self-Confidence	1.76	0.55	1.17	0.64	1.18	0.42	1.17	0.64
Intellectual Curiosity	1.62	0.89	1.10	1.00	1.24	0.54	1.10	1.00
Initiative	1.71	0.64	1.03	0.65	0.95	0.31	1.03	0.65
Social Orientation	1.12	0.95	0.89	0.97	1.07	0.56	0.89	0.97
Sincerity	1.29	0.77	0.82	0.67	0.98	0.51	0.82	0.67
Empathy	1.09	0.93	0.72	0.93	0.78	0.64	0.72	0.93
Innovation	0.99	0.93	0.71	1.01	0.83	0.52	0.71	1.01
Optimism	1.24	0.67	0.70	0.80	0.86	0.41	0.70	0.80
Leadership Orientation	0.95	0.95	0.66	1.02	0.87	0.59	0.66	1.02
Tolerance For Ambiguity	1.07	0.93	0.64	0.87	0.82	0.36	0.64	0.87
Humility	0.82	0.64	0.31	0.62	0.51	0.32	0.31	0.62

Note. *N* = 125 (unit of analysis is occupation). Ratings were made on a 7-point scale ranging from -3 (very detrimental to job performance) to +3 (very beneficial to job performance), with a 0-point corresponding to "little or no impact on job performance". Work Styles are sorted in descending order of their mean importance of expert ratings across the 125 occupations sampled.

Reliability and Agreement of Work Style Ratings by Source

Next, we examined the "operational" interrater reliability and agreement estimates for ratings based on each source for each Work Style. By "operational" estimates, we mean the reliability and agreement estimates for the final composite ratings from that source (i.e., not the reliability of ratings from a single rater or single LLM that contributes to the final rating from that source).

For analyst and expert ratings, ICC-based reliability and agreement coefficients are provided. ICC(C,*k*) reflects the estimated reliability of the mean rating for a given Work Style dimension (across *k* raters of a given type, eight for analysts, five for experts). 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. ICC(A,*k*) reflects the estimated absolute agreement of the mean rating for a given Work Style dimension (across eight analysts, or five experts). The expert reliability and agreement estimates, previously presented in Table 3.2, are repeated in Table 5.2 for the purpose of direct comparison with LLM and analyst estimates.

For LLM ratings, given that the final rating for a Work Style was based on a composite of ratings from three LLMs, where each LLM generated three ratings each (stemming from three

independent runs of the final prompt through each LLM), we employed a generalizability (G) theory approach to estimate reliability and agreement. Adopting a G-theory approach allowed us to simultaneously account for two types of measurement error in LLM ratings: (a) error stemming from inconsistency in ratings across the three different LLMs (i.e., model-specific error), (b) error stemming from inconsistency in ratings across occasions on which a given model was run (i.e., run-specific error) (Cronbach et al., 1972).

The measurement design underlying the LLM ratings for each Work Style can be described as occupations (o) crossed with runs (r) nested within models (i.e., o x [r:m]). Occupations represent the target of measurement, and as such, variation in ratings attributable to the occupation main effect may be viewed as "true score variance" in the Classical Test Theory (CTT) reliability sense ("universe score variance" in G-theory terms). Models (i.e., LLMs) and runs (i.e., runs of the prompt through a given model) can be considered measurement facets that have the potential to contribute to error in the observed Work Style ratings for occupations. Based on G-theory, we can (a) provide reliability and agreement estimates for LLM ratings that account for model-based, run-based, and residual sources of measurement error, and (b) partition the relative contribution of the aforementioned sources of error to overall measurement error in ratings.

To estimate reliability and agreement coefficients for the LLM ratings, we first estimated the Restricted Maximum Likelihood (REML) variance components for each effect that could be estimated in the measurement design above, namely: (1) occupation main effect (i.e., σ^2_o = variation in ratings across the targets of measurement), (2) model main effect (i.e., σ^2_m = variation in ratings across LLMs), (3) run nested within model effect (i.e., $\sigma^2_{r:m}$ = variation in ratings across runs within model), (4) the occupation-by-model interaction effect (i.e., σ^2_{om} = variation attributable to the occupation x LLM interaction), and the (5) residual effects (i.e., $\sigma^2_{or:m,res}$ = variation attributable to the occupation x LLM x run interaction and unexplained variation in ratings). We then used the resulting raw variance components to estimate the reliability of LLM ratings using the following formula based on rules from G-theory for a o x (r:m) measurement design.

$$\text{Reliability Estimate for LLM WS Ratings} = \frac{\sigma_o^2}{\sigma_o^2 + \left[\frac{\sigma_{om}^2}{3} + \frac{\sigma_{r:m}^2}{9} + \frac{\sigma_{or:m,res}^2}{9}\right]}$$

$$\text{Agreement Estimate for LLM WS Ratings} = \frac{\sigma_o^2}{\sigma_o^2 + \left[\frac{\sigma_m^2}{3} + \frac{\sigma_{om}^2}{3} + \frac{\sigma_{r:m}^2}{9} + \frac{\sigma_{or:m,res}^2}{9}\right]}$$

On average across Work Styles, the reliability of the LLM ratings was very high (.98), indicating that, on average, 98% of the observed variance in Work Style ratings (across occupations) was attributable to true (universe) score variance, while only 2% was attributable to measurement error variance. The average agreement estimate was only slightly lower, at .93. Stated differently, if we were to randomly sample another three LLMs and run our prompt through them three times each and take the average of the resulting ratings, we would expect the correlation between our final ratings and the new composite ratings to be approximately .98 on average, across Work Styles.

Also, of note in Table 5.2, LLM ratings appeared to be more reliable than ratings from either experts or analysts, and the reliability of expert ratings consistently exceeded that of analyst ratings, *despite* the final expert ratings being based on an average of five raters as opposed to eight for analysts. All else being equal, the more raters that are used to form a composite, the higher the interrater reliability of the resulting composite.

Table 5.2. Work Style Rating Reliability and Agreement Estimates by Source

Work Style	LLM		Analyst		Expert	
	G (Rel)	G (Abs)	ICC(C,8)	ICC(A,8)	ICC(C,5)	ICC(A,5)
Achievement Orientation	.98	.96	.84	.81	.85	.79
Adaptability	.98	.95	.72	.63	.84	.80
Attention To Detail	.97	.88	.82	.76	.80	.73
Cautiousness	.98	.93	.81	.78	.86	.83
Cooperation	.98	.93	.85	.76	.89	.83
Dependability	.95	.70	.67	.54	.57	.42
Empathy	.99	.96	.92	.89	.93	.91
Humility	.98	.96	.76	.68	.77	.74
Initiative	.98	.95	.65	.56	.76	.67
Innovation	.99	.97	.84	.77	.90	.88
Integrity	.98	.90	.75	.65	.85	.77
Intellectual Curiosity	.99	.98	.86	.75	.93	.89
Leadership Orientation	.99	.98	.90	.84	.91	.91
Optimism	.98	.95	.80	.71	.92	.89
Perseverance	.97	.90	.43	.36	.75	.72
Self-Confidence	.98	.95	.82	.72	.85	.72
Self-Control	.98	.91	.78	.69	.87	.85
Sincerity	.98	.96	.86	.81	.85	.77
Social Orientation	.99	.98	.88	.82	.92	.90
Stress Tolerance	.98	.93	.79	.69	.82	.79
Tolerance For Ambiguity	.97	.96	.62	.52	.84	.84
Average	.98	.93	.78	.70	.84	.79

Note. *N* = 125 (unit of analysis is occupation). Occupations are treated as targets of measurement.

Beyond the reliability and absolute agreement for ratings for the individual Work Style dimensions, we also examined the reliability (i.e., consistency) and absolute agreement (i.e., interchangeability) of the within-occupation Work Style profiles for LLM, analyst, and expert ratings. 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 125 occupations where Work Style dimensions serve as the target of measurement, that is, we address how consistent expert raters (or in the case of LLM ratings, LLM/run combinations) are in terms of their ordering of Work Style dimensions for any given occupation.

For analyst and expert ratings, ICC-based reliability and agreement coefficients are provided. ICC(C,*k*) reflects the estimated reliability of the mean Work Style profile for a given occupation (across *k* raters of a given type, eight for analysts, five for experts). Effectively, ICC(C,k) is comparable to the expected correlation one would expect to see between mean Work Style profiles provided by two randomly selected groups of *k* raters for a given occupation. ICC(A,*k*) reflects the estimated absolute agreement of the mean Work Style profiles for a given occupation (across eight analysts, or five experts). The expert reliability and agreement estimates, previously presented in Table 3.3, are repeated here for direct comparison with LLM and analyst estimates.

As we did when estimating the reliability and agreement coefficients for individual Work Styles based on LLM ratings, we adopted a G-theory approach that allowed us to simultaneously account for two types of measurement error to Work Style profiles: (a) error stemming from inconsistency in ratings across the three different LLMs (i.e., model-specific error), (b) error stemming from inconsistency in ratings across occasions on which a given model was run (i.e., run-specific error) (Cronbach et al., 1972).

The measurement design underlying the profile of LLM ratings for each occupation can be described as Work Styles (w) crossed with runs (r) nested within models (i.e., w x [r:m]). Work Styles represent the target of measurement, and as such, variation in ratings attributable to the Work Style main effect may be viewed as "true score variance" in CTT reliability sense ("universe score variance" in G-theory terms). Models (i.e., LLMs) and runs (i.e., runs of the prompt through a given model) can be considered measurement facets that have the potential to contribute to error in the observed Work Style profiles for an occupation. Based on G-theory, we can (a) provide reliability and agreement estimates for LLM Work Style profiles that account for model-based, run-based, and residual sources of measurement error, and (b) partition the relative contribution of the aforementioned sources of error to overall measurement error in ratings.

To estimate reliability and agreement coefficients for the profile of LLM ratings for each occupation, we first estimated the REML variance components for each effect that could be estimated in the measurement design above (for each occupation separately), namely: (1) Work Style main effect (i.e., σ^2_w = variation in ratings across the targets of measurement), (2) model main effect (i.e., σ^2_m = variation in ratings across LLMs), (3) run nested within model effect (i.e., $\sigma^2_{r:m}$ = variation in ratings across runs within model), (4) the Work Style-by-model interaction effect (i.e., σ^2_{wm} = variation attributable to the Work Style x LLM interaction), and (5) the residual effects (i.e., $\sigma^2_{or:m,res}$ = variation attributable to the Work Style x LLM x run interaction and unexplained variation in ratings). We then used the resulting raw variance component to estimate the reliability of LLM ratings using the following formula based on rules from G-theory for a w x (r:m) measurement design.

$$\text{Reliability Estimate for LLM WS Profiles} = \frac{\sigma^2_w}{\sigma^2_w + \left[\frac{\sigma^2_{wm}}{3} + \frac{\sigma^2_{r:m}}{9} + \frac{\sigma^2_{wr:m,res}}{9}\right]}$$

$$\text{Agreement Estimate for LLM WS Profiles} = \frac{\sigma^2_w}{\sigma^2_w + \left[\frac{\sigma^2_m}{3} + \frac{\sigma^2_{wm}}{3} + \frac{\sigma^2_{r:m}}{9} + \frac{\sigma^2_{wr:m,res}}{9}\right]}$$

Table 5.3 shows the results of our analysis. Recall, reliability, and agreement were estimated for each occupation, so the statistics in Table 5.3 reflect summaries of estimates across 125 occupations. On average, the operational reliability of the LLM-based Work Style profiles was very high (.98), indicating that, on average, 98% of the observed variance across Work Style ratings within an occupation was attributable to true (universe) score variance, while only 2% was attributable to measurement error. The average agreement estimate was only slightly lower, at .95. In comparison, on average, the operational reliability of analyst-based and expert-based Work Style profiles was .85 and .86, respectively, and on average, the operational agreement coefficients for analyst-based and expert-based Work Style profiles were .77 and .83, respectively. Thus, from both an operational reliability and agreement perspective, LLM-based Work Style profiles appear to offer an advantage over both analyst and expert ratings.

Table 5.3. Reliability and Agreement Estimates for Within-Occupation Work Style Profiles by Source

Statistic	LLM		Analyst		Expert	
	G(Rel)	G(Abs)	ICC(C,8)	ICC(A,8)	ICC(C,5)	ICC(A,5)
Mean	.98	.95	.85	.77	.86	.83
SD	.01	.02	.06	.09	.06	.08
Min	.94	.83	.56	.48	.62	.51
Max	1.00	.99	.93	.92	.94	.93
Percentiles						
5	.97	.91	.72	.60	.74	.66
25	.98	.94	.82	.74	.83	.79
50	.98	.96	.86	.80	.88	.85
75	.99	.97	.89	.84	.91	.89
95	.99	.98	.92	.87	.93	.91

Note. Separate G-coefficient and ICC estimates were calculated for each occupation, where the *n* for each estimate was 21 (i.e., the number of Work Styles for an occupation). The reported statistics are summaries of G-coefficients and ICC estimates across 125 occupations.

Comparison of Work Style Correlation Matrices Across Sources

Next, we evaluated the extent of differences in Work Style rating intercorrelations among LLM, analyst, and expert ratings. Appendix D provides Work Style correlation matrices based on calibrated LLM ratings and analyst ratings (see Tables E.1 and E.2, respectively).[10] The Work Style correlation matrix based on expert ratings was presented earlier in Table 3.4.

To summarize the extent of differences between the correlation matrix for each source of ratings, we first calculated the difference between the 120 corresponding correlations in each pair of correlation matrices (e.g., Achievement Orientation-Adaptability correlation based on LLM ratings – Achievement Orientation-Adaptability correlation based on expert ratings). We then calculated descriptive statistics summarizing correlation differences across the 120 Work Style pairs and presented them in Table 5.4.

[10] For these analyses, and all subsequent analyses our focus is only on the calibrated LLM ratings as those represent the draft final set of Work Style ratings being evaluated for use.

Table 5.4. Summary of Work Style Intercorrelation Differences between Sources

Statistic	LLM-Expert	Analyst-Expert	LLM-Analyst
Mean	.05	.05	.00
SD	.10	.15	.13
Min	-.25	-.27	-.40
Max	.31	.55	.35
Percentiles			
5	-.10	-.17	-.23
25	-.01	-.05	-.07
50	.05	.02	.01
75	.12	.14	.09
95	.21	.33	.19

Note. Cell values reflect summaries of pairwise differences between elements of the given sources' correlation matrices (e.g., descriptive statistics for pairwise differences between correlations for the 120 Work Styles pairs in the LLM Work Style correlation matrix and expert Work Style correlation matrix).

As shown in Table 5.4, on average differences between LLM and expert Work Style correlations tended to be small (*M* = .05) and on par with differences between analyst and expert Work Style correlations (*M* = .05). That being said, LLM Work Style ratings showed an advantage over analyst Work Style ratings in that they less variability in their difference with expert Work Style correlations (*SD* = .10, 5th percentile = -.10, and 95th percentile = .21) compared to variability in differences between analyst and expert correlations (*SD* = .15, 5th percentile -.17, 95th percentile. .33).

Convergence of Work Style Ratings Across Sources

Next, we evaluated the set of LLM, analyst, and expert ratings through a multitrait-multimethod (MTMM) lens to evaluate general patterns of convergence and discrimination among ratings from these sources. The purpose of doing so was to evaluate evidence of convergent and discriminant validity (Campbell & Fiske, 1959). Table 5.5 provides a summary of different types of MTMM correlations within and between the three rating sources.

Convergent validity for LLM-based Work Style ratings is indicated by high correlations among ratings for the same Work Styles based on different sources (i.e., monotrait-heteromethod correlations). As shown in Table 5.5, we see strong evidence of convergence between LLM and expert ratings. The average monotrait-heteromethod correlation among LLM and expert ratings was .84 (*SD* = .08), whereas among analyst and expert ratings it was .76 (*SD* = .10). Thus, from a convergent validity perspective, LLM ratings appeared more favorable than analyst ratings in terms of convergence with experts.

Discriminant validity and freedom from "common method" variance are indicated by relatively low correlations among different Work Styles rated by the same source (i.e., heterotrait-monomethod correlations) *and* that do not clearly exceed the correlations among different Work Style dimensions rated by different sources (i.e., heterotrait-heteromethod correlations). Again, focusing on the results in Table 5.5, we observe a pattern of evidence supporting discriminant validity, with minimal concern for common method variance in any of the rating sources. Specifically, the average heterotrait-monomethod correlation among LLM ratings was .40. Among analyst ratings, it was also .40, and among expert ratings, it was .35 (all clearly lower than the monotrait-heteromethod correlations summarized above). In comparison, the average

heterotrait-heteromethod correlation was .38 between LLM and expert ratings, .37 between analyst and expert ratings, and .40 between LLM and analyst, all of which are comparable to the average heterotrait-monomethod correlations.

Table 5.5. MTMM Correlation Summary

Correlation Type	*M*	*SD*	*Min*	*Max*	Percentiles				
					5	25	50	75	95
Monotrait-heteromethod (LLM, Expert)	.84	.08	.66	.95	.69	.81	.84	.88	.94
Monotrait-heteromethod (Analyst, Expert)	.76	.10	.54	.89	.54	.73	.78	.83	.88
Monotrait-heteromethod (LLM, Analyst)	.77	.10	.55	.91	.59	.71	.77	.85	.87
Heterotrait-monomethod (LLM)	.40	.27	-.43	.88	-.08	.23	.44	.60	.78
Heterotrait-monomethod (Analyst)	.40	.25	-.32	.87	-.04	.23	.44	.56	.78
Heterotrait-monomethod (Expert)	.35	.29	-.37	.89	-.16	.13	.41	.58	.77
Heterotrait-heteromethod (LLM,Expert)	.38	.29	-.46	.95	-.13	.17	.42	.59	.81
Heterotrait-heteromethod (Analyst, Expert)	.37	.27	-.41	.89	-.12	.18	.41	.56	.76
Heterotrait-heteromethod (LLM, Analyst)	.40	.26	-.42	.91	-.07	.24	.42	.57	.77

As a follow-up to the MTMM-focused analyses above, we examined convergence between LLM ratings and expert ratings more closely – specifically, for each individual Work Style. Table 5.6 presents two sets of correlations. One set is based on the full sample of 125 occupations for which expert and analyst ratings were gathered. The other set is based on a reduced set of occupations, where those occupations used as “few-shot” examples in the LLM prompt to generate LLM ratings for a given Work Style were removed. As noted at the end of Step 4, the reason for estimating correlations on this reduced sample was to determine how much LLM-expert correlations may be inflated or due to chance, given that expert ratings were ultimately used to finalize the set of occupations used as few-shot examples in LLM prompts for each Work Style.

A review of Table 5.6 reveals that on average, LLM ratings correlated very highly with experts (.84). This average only dropped slightly once occupations used as few-shot examples in LLM prompts were removed from consideration (.80). Furthermore, LLM-expert correlations tended to be higher on average (.84) then analyst-expert correlations (.76) – again, suggesting a performance advantage of LLMs over analysts (though this advantage was smaller based on the reduced sample: .80 vs. .75).

If we frame the LLM-expert correlations as criterion-related validity estimates for LLM ratings (treating expert ratings as the criterion), it is important to note that the correlations are attenuated by measurement error (unreliability) in the expert ratings. So, it is reasonable to consider what these correlations would be if they were corrected for unreliability in the criterion. These correlations are provided in Table 5.7 along with the uncorrected correlations for comparison.

A review of Table 5.7 reveals that after accounting for unreliability in expert ratings, the LLM-expert correlations are clearly higher on average (.84 vs. 91 in the full sample, .80 vs. 87 in the reduced sample). The largest improvements appeared to be for Dependability, which had the lowest uncorrected correlation with expert ratings (.66), but improved to .87 after accounting for error in the expert ratings. The pattern of findings was not surprising, as noted earlier, given that there was little variation in ratings of Dependability across experts. Under these conditions, there is a tendency for estimates of reliability and validity to be depressed.

Table 5.6. Between-Occupation Work Style Correlations

Work Style	Full Sample				Reduced Sample*			
	LLM-Expert	Analyst-Expert	Analyst-LLM		*n*	LLM-Expert	Analyst-Expert	Analyst-LLM
Leadership Orientation	.95	.89	.87		94	.93	.90	.87
Social Orientation	.94	.81	.87		90	.92	.80	.87
Innovation	.93	.83	.85		90	.91	.82	.86
Empathy	.93	.88	.91		90	.90	.86	.91
Intellectual Curiosity	.93	.88	.86		89	.93	.90	.87
Attention To Detail	.88	.77	.84		102	.83	.76	.86
Cooperation	.88	.83	.85		98	.87	.84	.88
Tolerance For Ambiguity	.87	.75	.71		92	.85	.76	.71
Optimism	.87	.84	.84		96	.85	.84	.83
Self-Confidence	.86	.73	.74		100	.88	.75	.83
Achievement Orientation	.85	.78	.80		97	.81	.79	.83
Cautiousness	.84	.80	.76		94	.77	.74	.70
Stress Tolerance	.84	.78	.77		98	.79	.76	.78
Self-Control	.83	.81	.76		96	.81	.83	.82
Integrity	.82	.75	.73		100	.81	.77	.80
Sincerity	.81	.76	.83		97	.77	.66	.85
Adaptability	.78	.71	.68		96	.73	.77	.74
Perseverance	.75	.54	.55		99	.68	.49	.54
Initiative	.73	.64	.68		98	.71	.64	.74
Humility	.69	.54	.59		97	.51	.59	.68
Dependability	.66	.59	.66		108	.47	.49	.63
Average	.84	.76	.77			.80	.75	.79

Note. Full Sample *N* = 125 for the full sample (unit of analysis is occupation). *Reduced Sample *n* = Number of occupations on which the given correlations are based, removing those occupations that were used as few-shot examples in the prompts used for LLM ratings for the given Work Style. Work Styles are sorted in descending order based on the magnitude of correlation between LLM and expert ratings.

Table 5.7. Comparison Between-Occupation LLM-Expert Work Style Correlations Corrected and Uncorrected for Unreliability in Expert Ratings

Work Style	Full Sample			Reduced Sample*		
	r_U	r_C		*n*	r_U	r_C
Leadership Orientation	.95	.99		94	.93	.98
Social Orientation	.94	.98		90	.92	.96
Innovation	.93	.98		90	.91	.95
Empathy	.93	.96		90	.90	.93
Intellectual Curiosity	.93	.96		89	.93	.96
Attention To Detail	.88	.98		102	.83	.93
Cooperation	.88	.93		98	.87	.92
Tolerance For Ambiguity	.87	.95		92	.85	.92
Optimism	.87	.91		96	.85	.88
Self-Confidence	.86	.93		100	.88	.96
Achievement Orientation	.85	.91		97	.81	.88
Cautiousness	.84	.91		94	.77	.83
Stress Tolerance	.84	.92		98	.79	.87
Self-Control	.83	.89		96	.81	.87
Integrity	.82	.89		100	.81	.88
Sincerity	.81	.88		97	.77	.83
Adaptability	.78	.85		96	.73	.80
Perseverance	.75	.87		99	.68	.78
Initiative	.73	.84		98	.71	.82
Humility	.69	.78		97	.51	.58
Dependability	.66	.87		108	.47	.63
Average	.84	.91			.80	.87

Note. Full Sample *N* = 125 for the full sample (unit of analysis is occupation). *Reduced Sample *n* = = Number of occupations on which the given correlations are based, removing those occupations that were used as few-shot examples in the prompts used for LLM ratings for the given Work Style. r_u = Uncorrected correlation between LLM and expert ratings. r_c = Correlation between correlation between LLM and expert ratings corrected for unreliability in expert ratings (using ICC[C,5] values for expert ratings from Table 5.2)

In addition to examining between-occupation correlations, we also examined within-occupation correlations of the Work Style profile. Unlike the correlations summarized in Tables 5.5 through 5.7 above, the unit of analysis for these correlations is Work Styles, which means the "sample size" when calculating them is 21 (i.e., the 21 Work Styles), and a separate correlation is calculated for each occupation. Essentially, these correlations help answer the question of whether Work Styles are consistently rated or ordered across different sources for a given occupation. Table 5.8 provides a summary of these within-occupation correlations between LLM and expert-based Work Style profiles, analyst and expert-based Work Style profiles, and analyst and LLM-based Work Style profiles.

As shown in Table 5.8, on average, LLM based Work Style profiles exhibited high correlations with expert-based Work Style profiles (*M* = 89, *SD* = .06). Those correlations tended to be higher than the correlations between analyst and expert-based Work Style profiles (*M* = .81, *SD*

= .09). Thus, from both a between- and within-occupation perspective, LLM ratings appear to be converging well with experts and offering an advantage over analyst ratings.

Table 5.8. Within-Occupation Work Style Profile Correlation Summary

Statistic	LLM-Expert	Analyst-Expert	Analyst-LLM
Mean	.89	.81	.82
SD	.06	.09	.09
Min	.67	.55	.50
Max	.97	.95	.95
Percentiles			
5	.75	.63	.66
25	.86	.76	.77
50	.90	.82	.84
75	.93	.88	.88
95	.96	.92	.91

Note. Separate Work Style profile correlations were calculated for each occupation, where the *n* for each correlation was 21 (i.e., the number of Work Styles in an occupation's profile). The reported statistics are summaries of correlations across 125 occupations.

Within-Occupation Differences in Work Style Ratings Across Sources

Although the correlations reported above provide insight into the similarity between LLM-based Work Style ratings and expert ratings from a correlational (rank-order) perspective, they do not address alignment/misalignment in absolute terms. As such, we evaluated the extent of within-occupation differences across LLM and expert ratings and we also examined whether those differences tended to vary by job zone and job family of the occupation.[11] Table 5.9 provides a summary of within-occupation differences between LLM and expert ratings.

Table 5.9 helps reinforce the relatively close correspondence between LLM and expert ratings. Specifically, for any given Work Style, the vast majority of LLM ratings fell within half a rating scale point of expert ratings (82.0% of occupations on average, across Work Styles). In contrast, for any given Work Style, relatively few LLM ratings overestimated experts by .5 a scale point or more (10.6% of occupations on average, across Work Styles), and relatively few LLM ratings underestimated experts by .5 a scale point or more (7.3% of occupations on average, across Work Styles). Furthermore, for any given Work Style, very few LLM ratings overestimated experts by 1 scale point or more (1.2% of occupations on average, across Work Styles), and even fewer LLM ratings underestimated experts by 1 scale point or more (0.3% of occupations on average, across Work Styles).

[11] For the sake of parsimony, our analyses in this section only focus on LLM-expert differences rather than analyst-expert or LLM-analyst differences, as the focal method of interest was the LLM ratings with expert ratings serving as the gold standard. By this point, the project team and Center were aligned with moving forward with LLM-based ratings as the method of choice.

Table 5.9. Within-Occupation Differences between LLM and Expert Work Style Ratings

Work Style	*M*	*SD*	*Min*	*Max*	% of Occupations where LLM-Expert Difference…				
					≥ +1.0	≥ 0.5 and < 1.0	> -0.5 and < 0.5	≤ -0.5 and > -1.0	≤ -1.0
Achievement Orientation	0.00	0.37	-0.96	0.94	0.0	8.8	86.4	4.8	0.0
Adaptability	0.00	0.48	-1.04	1.23	3.2	10.4	71.2	14.4	0.8
Attention To Detail	0.00	0.29	-1.36	0.91	0.0	6.4	92.0	0.8	0.8
Cautiousness	0.00	0.46	-2.14	1.28	1.6	11.2	79.2	7.2	0.8
Cooperation	0.00	0.39	-0.78	1.30	1.6	8.8	81.6	8.0	0.0
Dependability	0.00	0.29	-0.90	0.67	0.0	2.4	92.8	4.8	0.0
Empathy	0.00	0.36	-0.68	1.14	0.8	11.2	84.0	4.0	0.0
Humility	0.00	0.49	-1.40	1.39	4.0	13.6	72.8	8.0	1.6
Initiative	0.00	0.48	-0.91	1.21	3.2	12.0	69.6	15.2	0.0
Innovation	0.00	0.38	-0.91	1.07	1.6	12.0	80.8	5.6	0.0
Integrity	0.00	0.42	-0.82	1.46	1.6	8.0	81.6	8.8	0.0
Intellectual Curiosity	0.00	0.38	-0.78	1.17	0.8	6.4	83.2	9.6	0.0
Leadership Orientation	0.00	0.33	-1.00	0.98	0.0	8.8	87.2	4.0	0.0
Optimism	0.00	0.41	-1.01	2.10	0.8	11.2	81.6	5.6	0.8
Perseverance	0.00	0.39	-1.07	1.02	0.8	9.6	83.2	5.6	0.8
Self-Confidence	0.00	0.34	-0.62	0.90	0.0	4.8	89.6	5.6	0.0
Self-Control	0.00	0.42	-0.75	1.10	1.6	12.8	75.2	10.4	0.0
Sincerity	0.00	0.41	-0.79	1.85	1.6	8.8	84.0	5.6	0.0
Social Orientation	0.00	0.33	-0.60	0.80	0.0	11.2	84.0	4.8	0.0
Stress Tolerance	0.00	0.37	-0.57	1.46	1.6	8.8	84.8	4.8	0.0
Tolerance For Ambiguity	0.00	0.44	-1.09	1.27	0.8	11.2	77.6	9.6	0.8
Average	0.00	0.39	-0.96	1.20	1.2	9.4	82.0	7.0	0.3

Note. N = 125. For purposes of calculation, expert ratings were subtracted from LLM ratings, so positive differences indicate the LLM ratings overestimated expert ratings, and negative differences indicate LLM ratings underestimated expert ratings.

Tables 5.10 and 5.11 provide mean within-occupation differences between LLM and expert ratings by job zone and job family, respectively. On average, for any given Work Style, LLM-expert rating differences did not appear to vary significantly by job zone, ranging from 2 (occupations requiring some education or preparation) to 5 (occupations requiring extensive preparation). LLM ratings tended to underestimate expert ratings slightly for job zone 1 (occupations where little or no education/preparation is needed), but we suggest caution in overinterpreting these results, given that the effect appears small and is only based on a sample

of eight occupations for which LLM and expert ratings were available. A review of the results in Table 5.11 also reveals several clear trends in LLM-expert differences across job families. On average, for any given Work Style, LLM-expert differences were in the range of -.21 (Building and Grounds Cleaning Maintenance) to .18 (Life, Physical, and Social Science). However, we again suggest caution in overinterpreting the results, given that the effects appear to be small; the samples of occupations available for each job family on which these results were based were very small.

Table 5.10. Within-Occupation Differences between LLM and Expert Work Style Ratings by Job Zone

Work Style	Job Zone				
	1	2	3	4	5
Achievement Orientation	-0.17	-0.09	0.08	0.03	0.07
Adaptability	-0.07	0.04	0.10	0.01	-0.17
Attention To Detail	-0.15	-0.03	0.04	-0.03	0.10
Cautiousness	-0.27	-0.09	0.09	0.07	0.04
Cooperation	-0.16	0.04	-0.11	-0.02	0.16
Dependability	-0.37	-0.06	0.03	0.13	0.03
Empathy	-0.09	-0.06	-0.07	0.00	0.22
Humility	-0.22	-0.09	-0.02	0.09	0.14
Initiative	-0.34	-0.11	0.00	0.18	0.07
Innovation	-0.27	-0.01	0.07	0.06	-0.05
Integrity	-0.30	0.06	-0.04	0.05	-0.02
Intellectual Curiosity	-0.10	-0.01	0.02	-0.01	0.05
Leadership Orientation	0.09	-0.05	0.01	0.02	0.01
Optimism	0.39	0.06	-0.03	-0.10	-0.08
Perseverance	-0.40	0.07	-0.04	0.03	0.04
Self-Confidence	-0.20	0.00	0.08	0.04	-0.08
Self-Control	-0.26	0.03	0.01	0.02	0.01
Sincerity	-0.16	-0.01	0.01	0.00	0.07
Social Orientation	-0.06	-0.02	-0.07	-0.03	0.20
Stress Tolerance	-0.15	0.08	-0.01	-0.13	0.10
Tolerance For Ambiguity	-0.21	0.07	-0.12	-0.01	0.12
Average	-0.17	-0.01	0.00	0.02	0.05
n	8	38	27	30	22

Note. n = Number of occupations analyzed for the given job zone. For purposes of calculation, expert ratings were subtracted from LLM ratings, so positive differences indicate the LLM ratings overestimated expert ratings, and negative differences indicate LLM ratings underestimated expert ratings.

Table 5.11. Within-Occupation Differences in Work Style Ratings Across Sources by Job Family

Work Style	Job Family																				
	Architecture and Engineering	Arts, Design, Entertainment, Sports, and Media	Building and Grounds Cleaning and Maintenance	Business and Financial Operations	Community and Social Service	Computer and Mathematical	Construction and Extraction	Educational Instruction and Library	Farming, Fishing, and Forestry	Food Preparation and Serving Related	Healthcare Practitioners and Technical	Healthcare Support	Installation, Maintenance, and Repair	Legal	Life, Physical, and Social Science	Management	Office and Administrative Support	Personal Care and Service	Production	Protective Service	Sales and Related	Transportation and Material Moving
Achievement Orientation	0.05	0.02	-0.29	-0.07	0.46	-0.02	-0.39	-0.18	0.06	-0.02	0.23	-0.14	-0.04	0.22	0.02	0.05	-0.01	0.06	0.06	-0.12	-0.24	0.16
Adaptability	-0.35	-0.01	-0.26	0.00	0.32	-0.21	-0.50	-0.25	-0.09	0.03	0.33	0.10	0.11	-0.38	0.24	0.27	0.24	0.00	-0.02	-0.18	0.24	0.13
Attention To Detail	-0.21	-0.28	-0.39	-0.01	0.61	0.02	-0.03	-0.08	-0.04	-0.04	0.20	0.69	0.16	0.18	-0.07	0.00	0.04	0.05	-0.02	-0.13	0.12	-0.08
Cautiousness	0.03	0.06	-0.21	0.02	-0.10	0.21	0.11	0.10	-0.47	-0.15	-0.19	-0.08	0.19	0.14	0.22	0.15	0.36	-0.37	-0.19	0.18	-0.11	-0.11
Cooperation	-0.10	0.05	-0.45	-0.17	-0.22	-0.27	-0.11	0.01	0.08	-0.03	0.24	-0.22	0.00	-0.28	0.46	-0.02	-0.20	0.32	0.19	-0.29	-0.17	0.03
Dependability	0.24	-0.02	-0.17	-0.02	-0.08	0.05	-0.42	0.10	-0.06	-0.17	0.14	-0.20	0.02	0.18	0.11	0.05	-0.08	-0.01	0.00	-0.06	-0.01	-0.06
Empathy	-0.12	0.26	-0.34	-0.02	0.01	-0.11	0.02	0.52	-0.20	-0.05	0.02	0.19	-0.17	-0.15	0.27	0.01	-0.09	-0.12	-0.04	0.18	-0.15	-0.13
Humility	-0.22	0.38	-0.20	0.12	-0.18	0.01	-0.19	0.15	0.20	0.19	0.24	-0.18	-0.28	0.32	0.29	0.12	-0.09	0.35	-0.29	-0.14	-0.41	-0.15
Initiative	-0.12	0.48	-0.25	0.16	0.72	-0.10	0.15	-0.15	-0.41	-0.54	0.26	-0.81	0.01	0.09	0.32	0.25	-0.30	-0.08	-0.22	-0.03	-0.24	0.03
Innovation	-0.07	0.05	0.09	0.00	0.29	0.04	0.05	-0.28	-0.33	-0.18	0.23	-0.11	0.46	0.13	0.14	0.10	-0.33	0.25	-0.26	-0.11	0.06	-0.09
Integrity	0.02	-0.17	-0.69	-0.16	-0.01	0.03	-0.50	-0.05	0.03	0.00	0.16	0.37	0.12	0.03	0.24	0.00	-0.04	0.15	0.23	0.00	-0.14	0.01
Intellectual Curiosity	-0.22	-0.06	0.20	0.07	-0.09	-0.21	-0.16	-0.01	-0.16	-0.29	0.03	-0.11	0.23	0.60	0.08	0.35	0.02	0.10	-0.09	-0.22	-0.20	0.00
Leadership Orientation	0.20	-0.15	-0.10	0.08	0.06	-0.11	-0.20	-0.18	0.10	0.04	0.16	-0.52	-0.17	-0.21	0.22	0.10	-0.05	0.10	0.15	-0.11	-0.16	-0.09
Optimism	-0.15	0.67	0.35	-0.25	0.12	-0.18	0.06	0.00	0.07	0.48	-0.06	-0.25	-0.05	-0.20	0.24	0.00	-0.50	0.34	-0.11	-0.09	0.12	0.01
Perseverance	0.27	-0.28	-0.45	-0.01	-0.46	-0.09	0.02	-0.15	0.40	-0.21	0.29	0.30	-0.03	0.21	0.12	0.08	-0.15	-0.04	-0.02	0.00	-0.21	-0.12
Self-Confidence	0.38	-0.08	-0.20	-0.11	-0.28	-0.11	-0.06	0.08	-0.34	0.13	0.23	-0.13	0.13	0.10	-0.12	0.02	-0.17	0.08	-0.13	-0.17	-0.19	0.18
Self-Control	0.05	-0.22	-0.27	0.05	-0.45	0.03	0.11	-0.26	0.51	-0.50	0.30	-0.13	0.06	0.30	0.08	0.05	0.09	0.14	-0.10	-0.03	-0.42	-0.03
Sincerity	-0.18	0.19	-0.04	-0.10	0.05	-0.27	-0.04	0.04	0.00	0.02	0.15	-0.36	-0.18	-0.01	0.27	-0.01	-0.02	0.53	-0.15	0.30	-0.09	-0.02
Social Orientation	-0.25	0.01	-0.27	0.06	-0.07	0.26	-0.17	0.33	-0.37	-0.33	0.31	-0.14	-0.12	-0.13	0.18	-0.17	-0.01	-0.06	0.14	-0.32	0.02	0.01
Stress Tolerance	-0.14	-0.04	-0.25	-0.04	0.09	-0.14	0.22	-0.14	0.07	-0.30	0.10	0.15	0.25	0.21	0.12	-0.10	-0.08	0.17	-0.02	-0.02	-0.36	0.12
Tolerance For Ambiguity	0.12	0.22	-0.28	-0.04	-0.19	0.16	0.05	0.22	-0.04	0.05	-0.05	0.37	-0.06	0.19	0.27	-0.10	-0.37	0.02	-0.07	-0.41	0.13	-0.14
Average	-0.04	0.05	-0.21	-0.02	0.03	-0.05	-0.09	-0.01	-0.05	-0.09	0.16	-0.06	0.03	0.07	0.18	0.06	-0.08	0.09	-0.05	-0.08	-0.12	-0.02
n	8	5	3	8	2	5	7	7	3	3	10	2	6	3	7	8	6	5	12	3	4	8

Note. n = Number of occupations analyzed for the given job family. For purposes of calculation, expert ratings were subtracted from LLM ratings, so positive differences indicate the LLM ratings overestimated expert ratings, and negative differences indicate LLM ratings underestimated expert ratings.

Step 6: Summary of LLM Work Style Ratings Across All Data-Level Occupations

To further evaluate and summarize the functioning of the best-best LLM rating method emerging from the previous steps, we applied it to all 891 active, data-level occupations in O*NET 29.3 (not simply the 125 sampled for evaluation against expert ratings in Step 5). Note that, in contrast to Step 5 where analyses were based on O*NET 29.0, we shifted to O*NET 29.3 here, as it was the latest available data at the time work on this step of the project was conducted. The LLM ratings presented here reflect ratings that were calibrated using the process described in Step 5. In the sections that follow, we summarize Work Style rating descriptives and distributions, reliability and agreement estimates, and intercorrelations. In addition to summarizing the statistics above, we also conducted additional analyses with the aim of (a) revisiting the provisional higher-order Work Styles structure offered in Phase 1 of the Work Styles revision effort (see Table 1 in this report), and (b) identifying distinguishing Work Styles for each occupation that would be part of the Work Styles data (along with final Work Style impact ratings) in the O*NET Database starting with the O*NET 30.1 release.

Work Style Rating Descriptives and Distributions

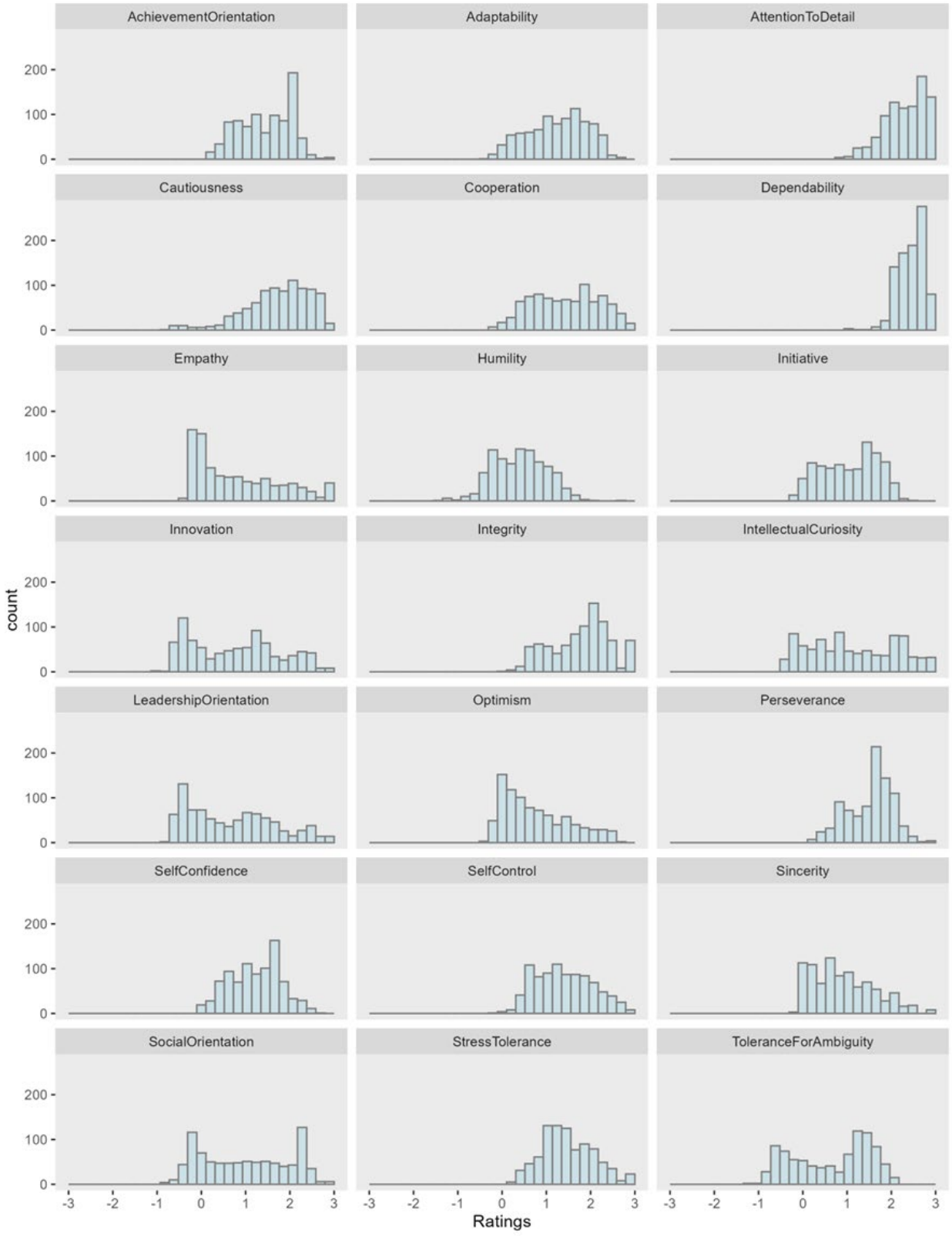
As shown in Table 6.1 and Figure 6.1, with the exception of Dependability and Attention to Detail, all Work Styles appear to differentiate among occupations, as evidenced by the level of standard deviation in the ratings and spread of the distributions. Dependability and Attention to Detail appear to be beneficial for the vast majority of occupations in O*NET, which seems intuitive, and is similar to the finding with expert ratings among the subset of 125 occupations sampled for earlier steps. Indeed, it is difficult to think of occupations where these Work Styles would not be beneficial. Although Work Styles in general appeared to help differentiate occupations, it is worth noting that, for the most part, ratings fell on the positive (beneficial) rather than negative (detrimental) side of the Work Styles impact rating scale. Based on the first phase of Work Styles work, this finding does not surprise us, as the Work Styles examined are generally positive traits. However, we believe that retaining the rating scale as given was important, as both experts and LLMs identified occupations where various Work Styles were rated as having a detrimental (even if small) impact on performance.

Table 6.1. Work Style Rating Descriptives across All Active Data-Level Occupations

Work Style	*M*	*SD*	*Min*	*Max*	Percentiles				
					5	25	50	75	95
Dependability	2.44	0.29	0.95	2.93	1.98	2.25	2.47	2.65	2.93
Attention To Detail	2.33	0.49	0.74	3.00	1.44	2.00	2.37	2.70	3.00
Integrity	1.79	0.66	0.08	2.93	0.68	1.29	1.89	2.24	2.93
Cautiousness	1.74	0.74	-0.79	3.00	0.40	1.32	1.86	2.28	2.72
Perseverance	1.51	0.50	0.20	3.00	0.60	1.12	1.64	1.80	2.19
Stress Tolerance	1.47	0.59	0.15	3.00	0.55	1.04	1.40	1.83	2.55
Achievement Orientation	1.45	0.59	0.20	2.91	0.49	0.95	1.55	2.01	2.21
Cooperation	1.44	0.75	-0.27	3.00	0.27	0.80	1.44	2.07	2.62
Self-Control	1.40	0.64	-0.17	3.00	0.46	0.85	1.37	1.91	2.50
Adaptability	1.25	0.67	-0.32	2.74	0.10	0.73	1.29	1.79	2.26
Self-Confidence	1.22	0.57	-0.10	2.69	0.29	0.78	1.26	1.66	2.16
Intellectual Curiosity	1.14	0.98	-0.50	3.00	-0.26	0.31	1.02	2.04	2.72
Initiative	1.06	0.63	-0.29	2.50	0.03	0.51	1.13	1.56	2.00
Social Orientation	0.99	0.98	-0.77	2.81	-0.35	0.02	0.99	1.92	2.40
Sincerity	0.91	0.69	-0.12	2.89	0.04	0.30	0.78	1.39	2.17
Empathy	0.80	0.95	-0.39	3.00	-0.18	-0.05	0.52	1.54	2.62
Innovation	0.76	1.00	-0.95	3.00	-0.59	-0.24	0.79	1.47	2.46
Optimism	0.76	0.75	-0.34	2.77	-0.13	0.15	0.59	1.32	2.22
Leadership Orientation	0.68	0.99	-0.81	3.00	-0.56	-0.22	0.59	1.42	2.47
Tolerance For Ambiguity	0.66	0.87	-1.29	2.08	-0.69	-0.23	0.91	1.43	1.81
Humility	0.40	0.59	-1.42	2.62	-0.44	-0.08	0.42	0.84	1.35

Note. *N* = 891 (unit of analysis is occupation). Ratings were made on a 7-point scale ranging from -3 (very detrimental to job performance) to +3 (very beneficial to job performance), with a 0-point corresponding to "little or no impact on job performance". Work Styles are sorted in descending order of their mean importance across the 891 occupations.

Figure 6.1. Work Style Rating Density Plots across All Active Data-Level Occupations



Work Style Ratings by Job Zone and Job Family

Next, we examined differences in Work Style ratings by job zone and job family (see Tables 6.2 and 6.3), respectively. For each Work Style, we also decomposed variance ratings across occupations to evaluate what percentage of observed variance was attributable to job zone and job family, respectively. This decomposition was accomplished by fitting a nested random effects model for each Work Style, where occupations were treated as nested within job zone, and occupations were treated as nested within job family (separate models for job zone and job family), and estimating REML variance components for job zone, job family, and residual sources of variance. The Work Styles in Tables 6.2 and 6.3 are presented in descending order of the percentage of variance in their ratings that is attributed to job zone and job family, respectively, to highlight those Work Styles where ratings tend to vary most across job zones and job families.

Starting with Table 6.2, on average, for any given Work Style, 43.1% of the variance in ratings across occupations was attributable to job zone. Work Styles varied greatly in terms of the percentage of variance in their ratings that could be attributed to job zone, with Intellectual Curiosity (83.0%) and Achievement Orientation (74.5%) falling at the top of the distribution, such that those Work Styles were rated as far more beneficial for occupations in higher job zones than lower job zones. In contrast, Cautiousness (7.8%) and Optimism (11.4%) fell at the bottom of the distribution, indicating that these Work Styles tended to vary little across job zones and generally received lower to middling ratings. From a career exploration perspective, this set of findings suggests that some Work Styles may matter more than others (i.e., those that are more differentiated across job zones) when considering occupations in different job zones when it comes to personality-related occupation fit.

With respect to job family, on average, for any given Work Style, 53.8% of the variance in ratings across occupations was attributable to job family. Work Styles varied greatly in terms of the percentage of variance in their ratings that could be attributed to job family, with Intellectual Curiosity (69.7%), Empathy (66.2%), Sincerity (65.4%), and Social Orientation (64.2%) falling at the top of the distribution. In the case of all four of these Work Styles, there was a clear split between job families where the given Work Style that had very high ratings (e.g., Life, Physical, and Social Science for Intellectual Curiosity; Community and Social Service for Empathy, Sincerity, and Social Orientation) and job families where the given Work Style had very low ratings (e.g. Construction and Extraction, and Food Preparation and Serving Related for Intellectual Curiosity; Construction and Extraction, and Food Preparation and Serving Related for Empathy, Sincerity, and Social Orientation). In contrast, Dependability (32.4%) ranked at the bottom of the percentage of variance due to job family distribution, where it tended to receive high ratings across all job families. Again, from a career exploration perspective, this set of findings suggests that some Work Styles may matter more than others when considering occupations in different job families in terms of personality-related occupation fit.

Table 6.2. Mean Work Style Ratings by Job Zone

Work Style	% Variance Due to Job Zone	Job Zone				
		1	2	3	4	5
Intellectual Curiosity	83.0	-0.14	0.18	1.07	1.82	2.39
Achievement Orientation	74.5	0.68	0.91	1.43	1.94	2.03
Tolerance For Ambiguity	69.4	-0.31	-0.11	0.57	1.33	1.51
Innovation	67.2	-0.27	-0.13	0.69	1.57	1.69
Initiative	64.3	0.31	0.53	1.03	1.58	1.57
Self-Confidence	57.1	0.49	0.80	1.22	1.61	1.63
Adaptability	56.3	0.45	0.73	1.29	1.73	1.70
Integrity	50.7	0.85	1.40	1.84	2.10	2.23
Leadership Orientation	46.6	-0.28	0.03	0.58	1.34	1.35
Perseverance	46.1	1.02	1.15	1.57	1.81	1.83
Sincerity	41.2	0.24	0.54	0.95	1.03	1.55
Social Orientation	34.5	0.10	0.52	0.94	1.31	1.73
Empathy	33.4	0.05	0.40	0.78	0.87	1.67
Cooperation	32.4	0.72	1.09	1.49	1.58	1.97
Attention To Detail	31.6	1.74	2.14	2.49	2.43	2.46
Self-Control	27.2	0.75	1.18	1.49	1.48	1.75
Dependability	26.2	2.12	2.35	2.51	2.49	2.51
Humility	24.6	-0.01	0.17	0.46	0.41	0.83
Stress Tolerance	19.4	0.99	1.28	1.54	1.56	1.73
Optimism	11.4	0.45	0.55	0.74	0.86	1.14
Cautiousness	7.8	1.22	1.71	1.90	1.68	1.73
Average	43.1	0.53	0.83	1.27	1.55	1.76
n		31	289	211	211	149

Note. *N* = 891 (unit of analysis is occupation). Ratings were made on a 7-point scale ranging from -3 (very detrimental to job performance) to +3 (very beneficial to job performance), with a 0-point corresponding to "little or no impact on job performance". Work Styles are sorted in descending order of the percentage of their variance across occupations that is due to the job zone.

Table 6.3. Mean Work Style Ratings by Job Family

Work Style	% Variance Due to Job Family	Job Family																					
		Architecture and Engineering	Arts, Design, Entertainment, Sports, and Media	Building and Grounds Cleaning and Maintenance	Business and Financial Operations	Community and Social Service	Computer and Mathematical	Construction and Extraction	Educational Instruction and Library	Farming, Fishing, and Forestry	Food Preparation and Serving Related	Healthcare Practitioners and Technical	Healthcare Support	Installation, Maintenance, and Repair	Legal	Life, Physical, and Social Science	Management	Office and Administrative Support	Personal Care and Service	Production	Protective Service	Sales and Related	Transportation and Material Moving
Intellectual Curiosity	69.7	2.07	1.42	0.20	1.60	1.39	2.33	0.09	2.14	0.34	0.08	1.84	0.47	0.87	1.88	2.38	1.51	0.16	0.37	0.33	0.69	0.70	0.23
Empathy	66.2	0.17	0.75	0.29	0.74	2.87	0.27	-0.08	1.92	0.08	0.83	1.99	2.08	0.11	0.85	0.61	1.22	0.72	1.59	-0.10	0.95	1.20	0.41
Sincerity	65.4	0.59	0.75	0.49	0.97	2.59	0.56	0.21	1.62	0.41	0.78	1.77	1.65	0.52	1.27	1.02	1.20	0.88	1.41	0.16	1.00	0.93	0.64
Social Orientation	64.2	0.52	1.42	0.40	1.32	2.40	0.58	-0.10	2.15	-0.11	1.35	1.66	1.70	0.22	1.12	0.88	1.88	1.08	1.83	-0.16	1.02	2.42	0.62
Innovation	62.3	1.86	1.93	0.19	0.89	1.23	1.97	-0.14	1.67	-0.18	0.23	1.04	0.07	0.48	0.38	1.72	1.40	-0.24	0.40	-0.04	0.10	0.74	-0.31
Optimism	60.3	0.26	1.08	0.68	0.58	2.24	0.26	0.17	1.60	0.35	1.33	1.30	1.61	0.29	0.46	0.49	1.20	0.67	1.59	0.01	0.73	1.70	0.64
Tolerance For Ambiguity	60.0	1.06	1.32	0.00	1.14	1.44	1.56	-0.36	1.39	0.18	0.14	1.09	0.39	0.16	1.08	1.41	1.40	-0.03	0.53	-0.43	0.94	0.91	0.06
Achievement Orientation	58.0	1.92	1.76	0.97	1.85	1.60	1.95	0.94	1.88	0.90	0.78	1.79	0.93	1.20	1.74	1.91	2.11	0.89	0.88	0.98	1.36	1.74	1.00
Humility	57.9	0.31	-0.19	0.33	0.37	1.57	0.38	0.00	0.95	0.04	0.72	0.95	1.18	0.10	0.61	0.66	0.40	0.50	0.96	-0.15	0.21	-0.43	0.25
Initiative	55.5	1.25	1.49	0.86	1.43	1.74	1.52	0.34	1.58	0.58	0.83	1.31	0.85	0.72	1.05	1.43	1.92	0.60	0.99	0.31	1.28	1.45	0.64
Cooperation	55.3	1.17	1.43	1.15	1.37	2.63	1.20	0.82	2.16	0.69	1.80	2.26	2.31	0.88	1.37	1.48	1.92	1.45	2.04	0.51	1.35	1.60	1.34
Integrity	54.8	1.91	1.21	1.34	2.34	2.40	2.06	1.14	2.04	1.39	1.21	2.28	2.05	1.49	2.75	2.10	2.33	1.82	1.64	0.99	2.69	1.68	1.71
Self-Confidence	53.2	1.37	1.69	0.73	1.63	1.60	1.35	0.76	1.50	0.82	0.78	1.55	0.81	0.92	1.68	1.30	1.92	0.73	1.11	0.57	1.66	1.83	1.06
Cautiousness	49.9	2.18	0.46	1.52	1.91	1.41	1.99	1.86	1.06	1.56	0.75	2.27	1.86	2.02	2.29	1.86	1.65	1.49	1.23	1.91	2.29	0.44	2.17
Attention To Detail	47.8	2.73	1.97	1.76	2.60	1.88	2.76	2.00	1.93	1.86	1.73	2.80	2.33	2.43	2.74	2.57	2.21	2.46	1.94	2.42	2.25	1.78	2.11
Adaptability	47.5	1.59	1.77	0.73	1.39	1.88	1.82	0.56	1.68	0.85	1.01	1.63	1.11	0.94	0.88	1.64	1.87	0.77	1.14	0.46	1.60	1.40	1.00
Leadership Orientation	46.6	0.89	0.92	0.58	1.17	1.33	0.63	-0.03	1.43	0.24	0.26	1.03	-0.14	0.00	0.98	0.93	2.50	-0.11	0.69	-0.30	1.41	1.05	0.50
Self-Control	46.4	1.05	1.33	1.00	1.56	2.12	1.02	0.93	1.69	1.10	1.42	2.09	1.73	1.02	2.03	1.21	1.92	1.27	1.67	0.79	2.39	1.54	1.52
Perseverance	41.0	1.67	1.63	1.29	1.74	1.86	1.77	1.36	1.67	1.54	0.91	1.83	1.39	1.54	1.81	1.71	1.92	0.94	1.14	1.00	1.93	1.90	1.29
Stress Tolerance	34.5	1.21	1.52	1.13	1.53	2.09	1.36	1.28	1.53	1.25	1.44	2.06	1.64	1.29	1.79	1.32	1.89	1.13	1.43	0.95	2.40	1.50	1.62
Dependability	32.4	2.46	2.07	2.31	2.50	2.64	2.45	2.32	2.45	2.33	2.25	2.68	2.57	2.44	2.64	2.37	2.66	2.49	2.37	2.25	2.72	2.21	2.54
Average	53.8	1.34	1.32	0.86	1.46	1.95	1.42	0.67	1.72	0.77	0.98	1.77	1.36	0.94	1.49	1.48	1.76	0.94	1.28	0.59	1.47	1.35	1.00
n		53	40	8	45	14	36	57	62	12	16	89	19	49	7	56	45	51	31	104	26	21	50

Note. *N* = 891 (unit of analysis is occupation). Ratings were made on a 7-point scale ranging from -3 (very detrimental to job performance) to +3 (very beneficial to job performance), with a 0-point corresponding to “little or no impact on job performance”. Work Styles are sorted in descending order of the percentage of their variance across occupations that is due to job family.

Reliability and Agreement of Work Style Ratings

As we did for the LLM ratings in Step 5, we adopted a generalizability (G) theory approach to estimate reliability and agreement for the LLM ratings for the full set of 891 data-level occupations evaluated here. As noted in Step 5, adopting a G-theory approach allowed us to simultaneously account for two types of measurement error in LLM ratings: (a) error stemming from inconsistency in ratings across the three different LLMs (i.e., model-specific error), and (b) error stemming from inconsistency in ratings across occasions on which a given model was run (i.e., run-specific error) (Cronbach et al., 1972).

Unlike Step 5, which provides "operational" estimates, that is, reliability and agreement estimates for the final composite LLM ratings, we also offer "single-model, single-run" estimates. That is the expected reliability and agreement of ratings produced by a single LLM, where our final prompt was only run once. We provide these estimates to illustrate the value of basing LLM ratings on multiple models and multiple runs, thereby reducing the contribution of model- and run-specific variance to final ratings.

Table 6.4 presents both operational (composite rating) and single-model-single-run reliability and agreement estimates. The operational reliability of the LLM ratings was very high (.98), indicating that on average, 98% of the observed variance in Work Style ratings (across occupations) was attributable to true (universe) score variance, while only 2% was attributable to measurement error. The average agreement estimate was only slightly lower, at .93. Stated differently, if we were to randomly sample another three LLMs and run our prompt through them three times each and take the average of the resulting ratings, we would expect the correlation between our final ratings and the new composite ratings to be approximately .98 on average, across Work Styles. Note that these findings are identical (on average) to the results presented for the LLM ratings based on the 125 occupations analyzed for Step 5.

With respect to the composite vs. single model-single run estimates, differences were substantial. Specifically, on average, for any given Work Style, the reliability of the LLM composite rating was .98 whereas on average, for any given Work Style, the expected reliability of ratings produced by a single model with a single run was .82 (a difference of .16). Additionally, on average, for any given Work Style, the agreement estimate for the LLM composite rating was .93 whereas on average, for any given Work Style, the agreement estimate for ratings produced by a single model with a single run was .74 (a difference of .19). This indicates that by basing a final LLM rating on an average of multiple models and multiple runs can greatly reduce the contribution of idiosyncratic model-specific and run-specific error to observed variance in ratings.

Table 6.4. Work Style Rating Reliability and Agreement for Work Style Dimensions

Work Style	Composite Rating			Single Model-Single Run	
	G (Rel)	G (Abs)		G (Rel)	G (Abs)
Achievement Orientation	.98	.96		.86	.81
Adaptability	.98	.96		.84	.79
Attention To Detail	.97	.86		.76	.59
Cautiousness	.98	.91		.83	.70
Cooperation	.98	.92		.83	.72
Dependability	.93	.60		.61	.30
Empathy	.99	.96		.92	.85
Humility	.95	.95		.70	.69
Initiative	.98	.96		.82	.78
Innovation	.99	.97		.89	.85
Integrity	.98	.90		.82	.67
Intellectual Curiosity	.99	.98		.93	.90
Leadership Orientation	.98	.97		.86	.84
Optimism	.98	.96		.83	.79
Perseverance	.97	.93		.79	.71
Self-Confidence	.98	.95		.86	.78
Self-Control	.97	.87		.80	.62
Sincerity	.98	.96		.85	.81
Social Orientation	.99	.98		.92	.90
Stress Tolerance	.98	.94		.82	.75
Tolerance For Ambiguity	.97	.95		.78	.74
Average	.98	.93		.82	.74

Note. *N* = 891 (unit of analysis is occupation). Occupations are treated as targets of measurement.

In addition to estimating the reliability and agreement of ratings for each Work Style, we also estimated the reliability and agreement of within-occupation Work Style profiles (i.e., consistency and absolute agreement in Work Style profiles for a given occupation across LLMs/runs) as we did in Step 5. Once again, we adopted a G-theory approach that allowed us to simultaneously account for two types of measurement error in LLM ratings: (a) error stemming from inconsistency in ratings across the three different LLMs (i.e., model-specific error), and (b) error stemming from inconsistency in ratings across occasions on which a given model was run (i.e., run-specific error) (Cronbach et al., 1972).

Again, unlike Step 5, which provides “operational” estimates, that is, reliability and agreement estimates for Work Style profiles based on final composite LLM ratings, we also provide “single-model, single-run” estimates. That is the expected reliability and agreement of Work Style profiles produced by a single LLM, where our final prompt was only run once for each Work Style. We again provide these estimates to illustrate the value of basing LLM Work Style profiles

on multiple models and runs, thereby reducing the contribution of model- and run-specific variance to final ratings.

Table 6.5 presents both operational (composite rating) and single-model-single-run reliability and agreement estimates for Work Style profiles. Recall, reliability, and agreement were estimated for each occupation, so the statistics in Table 6.5 reflect summaries of estimates across 891 occupations. On average, the operational reliability of the Work Style profiles was very high (.98), which means that on average, 98% of observed variance across Work Style ratings within an occupation was attributable to true (universe) score variance, and only 2% was attributable to measurement error. The average agreement estimate was only slightly lower, at .95.

With respect to the composite vs. single model-single run estimates, differences were once again substantial. Specifically, on average, for any given occupation, the reliability of the Work Style profiles based on the composite rating was .98 whereas on average, for any given occupation, the expected reliability of Work Style profiles produced by a single model with a single run was .84 (a difference of .16). Additionally, on average, for any given occupation, the agreement estimate for the Work Style profiles based on the composite rating was .95 whereas on average, for any given occupation, the agreement estimate for Work Style profiles produced by a single model with a single run was .79 (a difference of .16). This indicates that by basing a final LLM rating on an average of multiple models and multiple runs can greatly reduce the contribution of idiosyncratic model-specific and run-specific error to observed variance in Work Style profiles when profiling a given occupation.

Table 6.5. Reliability and Agreement for Within-Occupation Work Style Profiles

Statistic	Composite Rating			Single Model-Single Run	
	G(Rel)	G(Abs)		G(Rel)	G(Abs)
Mean	.98	.95		.84	.79
SD	.01	.02		.06	.07
Min	.93	.83		.59	.53
Max	.99	.99		.95	.94
Percentiles					
5	.96	.91		.74	.67
25	.98	.94		.81	.75
50	.98	.96		.85	.80
75	.99	.97		.88	.84
95	.99	.98		.92	.89

Note. Separate G-coefficients were estimated for each occupation, where the *n* for each estimate was 21 (i.e., the number of Work Styles for an occupation). The reported statistics are summaries of estimated G-coefficients across 891 occupations.

Work Style Intercorrelations

Next, we examined correlations among Work Styles to evaluate whether they were empirically distinct based on LLM ratings (i.e., they did not exhibit perfect or near perfect correlations). A review of Work Style correlations in Table 6.6 reveals only one pair of Work Styles exhibited correlations of .90 or greater in magnitude (Empathy-Sincerity, r = .93), which would be suggestive of empirical redundancy among the pair considered. That being said, there were a

small number of clusters of Work Styles that exhibited very high correlations of .87 or above in magnitude, those being:

- Empathy not only had very high correlations with Sincerity (.93), but also Optimism (.88), and Social Orientation (.87). Optimism and Social Orientation were also correlated .88.[12]
- Tolerance for Ambiguity had very high correlations with Initiative (.89), but also Adaptability (.87). Initiative and Adaptability were also correlated .88.
- Self-Control had a very high correlation with Stress Tolerance (.89)

It is worth noting, these we observed similar and very high correlations among expert rating presented in Table 3.4 earlier (e.g. Self-Control-Stress Tolerance, .89; Empathy-Optimism, .88; Optimism-Social Orientation, .87), so the findings here do necessary suggest LLM-based ratings were inflating correlations among these Work Styles in a way that deviated from experts, but rather confirms that these Work Styles simply covary extremely highly when occupations are used as the units of analysis. Out of all 210 pairs of Work Styles, 165 pairs (78.6%) had correlations of less than .70 in magnitude, and the average correlation was .45 (*SD* = .29, *Min* = -.42 [Cautiousness, Optimism, *Max* = .93 [Empathy, Sincerity]).

[12] Despite the high correlation between Empathy and Sincerity, we decided not to combine them based on a review of other data for these Work Styles, which suggested they were not completely redundant (e.g., differences in means and *SD*s across occupations, mean differences observed across job zones and job families). In other words, the correlation across occupations provides only one index of similarity between Work Styles, but when other forms of similarity are considered, these Work Styles appear to have some distinguishing value. As such, we erred on the side of keeping them separate rather than combining them or retaining only one of them.

Table 6.6. Work Style Rating Intercorrelations

Work Style		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Achievement Orientation	1.00																				
2	Adaptability	**.73**	1.00																			
3	Attention To Detail	**.31**	**.07**	1.00																		
4	Cautiousness	.05	-.06	**.68**	1.00																	
5	Cooperation	**.32**	**.62**	**-.11**	**-.12**	1.00																
6	Dependability	**.29**	**.36**	**.43**	**.61**	**.42**	1.00															
7	Empathy	**.28**	**.53**	**-.16**	**-.20**	**.87**	**.30**	1.00														
8	Humility	**.13**	**.35**	.06	**.11**	**.75**	**.40**	**.72**	1.00													
9	Initiative	**.81**	**.87**	.03	**-.15**	**.59**	**.30**	**.54**	**.34**	1.00												
10	Innovation	**.78**	**.71**	**.12**	**-.21**	**.30**	-.04	**.28**	**.17**	**.77**	1.00											
11	Integrity	**.54**	**.55**	**.39**	**.41**	**.50**	**.72**	**.45**	**.46**	**.56**	**.25**	1.00										
12	Intellectual Curiosity	**.86**	**.67**	**.41**	**.11**	**.31**	**.23**	**.26**	**.26**	**.72**	**.83**	**.52**	1.00									
13	Leadership Orientation	**.71**	**.73**	-.02	-.04	**.55**	**.41**	**.48**	**.30**	**.82**	**.55**	**.56**	**.54**	1.00								
14	Optimism	**.21**	**.52**	**-.41**	**-.42**	**.82**	**.17**	**.88**	**.57**	**.53**	**.27**	**.28**	**.11**	**.47**	1.00							
15	Perseverance	**.77**	**.74**	**.18**	**.15**	**.39**	**.47**	**.36**	**.17**	**.74**	**.54**	**.60**	**.61**	**.66**	**.33**	1.00						
16	Self-Confidence	**.79**	**.80**	.06	-.03	**.52**	**.38**	**.49**	**.20**	**.82**	**.60**	**.60**	**.62**	**.83**	**.49**	**.79**	1.00					
17	Self-Control	**.37**	**.59**	.05	**.18**	**.70**	**.66**	**.69**	**.49**	**.51**	**.09**	**.72**	**.24**	**.60**	**.61**	**.60**	**.67**	1.00				
18	Sincerity	**.35**	**.55**	-.02	-.04	**.86**	**.43**	**.93**	**.79**	**.57**	**.30**	**.60**	**.36**	**.51**	**.79**	**.43**	**.52**	**.73**	1.00			
19	Social Orientation	**.41**	**.62**	**-.24**	**-.37**	**.84**	**.21**	**.87**	**.56**	**.69**	**.41**	**.45**	**.32**	**.63**	**.88**	**.41**	**.65**	**.64**	**.81**	1.00		
20	Stress Tolerance	**.40**	**.61**	**.10**	**.28**	**.56**	**.69**	**.51**	**.33**	**.49**	**.10**	**.65**	**.26**	**.57**	**.44**	**.67**	**.66**	**.89**	**.55**	**.46**	1.00	
21	Tolerance For Ambiguity	**.80**	**.87**	**.09**	**-.12**	**.53**	**.23**	**.51**	**.34**	**.89**	**.81**	**.56**	**.80**	**.71**	**.46**	**.72**	**.79**	**.47**	**.54**	**.61**	**.46**	1.00

Note. N = 891 (unit of analysis is occupation). Correlations that are bolded are statically significant (p < .05, two-tailed).

Revisiting the Higher-Order Work Styles Structure

As noted in the introduction of this report, as part of an earlier part of the Work Styles revision effort, a preliminary, updated higher-order structure for the Work Styles taxonomy was proposed (see Putka et al., 2024). That higher-order structure was viewed as preliminary by the Center and project team, given that it would be revisited once updated Work Style ratings using the best-bet method that emerged from the current phase of research were identified and applied to all active data-level occupations. Doing so would allow us to examine the factor structure underlying the ratings using the complete set of active, data-level occupations as the basis for analysis, and suggest revisions to the higher-order structure that reflect the observed empirical structure, should such a structure be deemed meaningful (e.g., the factor solution exhibits elements of simple structure, the resulting factors are substantively meaningful).

As part of this step, we used the best-bet LLM method to generate Work Style ratings for all 891 active, data-level occupations in O*NET 29.3. We conducted a Principal Components Analysis (PCA) on the data with oblique (promax) rotation, and based on results of a parallel analysis, arrived at a four-component solution.[13] The resulting solution accounted for 87.6% of variance in the Work Style ratings, with 52.1% accounted for by the first factor, 15.9% by the second, 13.5% by the third, and 6.1% by the fourth. Table 6.7 provides the pattern matrix from the four-component PCA solution.

Table 6.7. Work Style Rating Principal Components Analysis Pattern Matrix

Work Style	Component			
	1	2	3	4
Innovation	.99	-.03	-.08	.36
Achievement Orientation	.95	.12	.13	-.07
Intellectual Curiosity	.93	-.06	.30	.26
Tolerance For Ambiguity	.87	-.17	-.05	-.01
Initiative	.82	-.18	-.12	-.12
Adaptability	.74	-.17	-.06	-.22
Self-Confidence	.71	.00	-.11	-.46
Perseverance	.68	.11	.08	-.48
Leadership Orientation	.63	-.08	-.11	-.40
Humility	-.10	-1.00	.27	.21
Sincerity	.09	-.90	.04	-.07
Empathy	.05	-.90	-.15	-.07
Cooperation	.09	-.85	-.07	-.12
Optimism	.05	-.74	-.44	-.17
Social Orientation	.26	-.69	-.33	-.13
Cautiousness	-.15	.10	.88	-.24
Attention To Detail	.22	.06	.88	.16
Dependability	-.05	-.24	.59	-.56
Integrity	.29	-.31	.47	-.37
Stress Tolerance	.12	-.16	.13	-.81
Self-Control	.04	-.43	.09	-.69

Note. N = 891 (unit of analysis is occupation).

[13] We conducted a parallel analysis using and found a critical eigenvalue to use in evaluating the number of components to extract (Horn, 1965). We also examined three- and five-component solutions to evaluate whether they offered a more interpretable solution than the four-component solution; however, we found that the four-component solution provided the cleanest interpretation.

As shown in Table 6.7, four substantively distinct components emerged from the analysis that we interpreted as follows:

- Component 1: Proactive and Growth Oriented
- Component 2: Interpersonally Oriented
- Component 3: Conscientious and Rule Oriented
- Component 4: Emotionally Resilient

Note that these components differ notably from the higher-order structure originally outlined in the development of the new set of Work Styles (see Putka et al., 2024). Table 6.8 provides a side-by-side comparison of the provisional higher-order structure identified in the first phase of the Work Styles revision effort (Putka et al, 2024) and the one based on the PCA above. One key reason for the differences observed here is that the provisional higher order structure identified by Putka et al (2024) was heavily informed by the body of past research on the factor structure underling indicators of *individual* personality (i.e., based on factor analytic work where *individuals*, rather than *occupations* were the units of analysis). In contrast, this was the first large scale work we know of that that looked at the factor structure of indicators personality-related requirements of occupations, where occupations, rather than individuals were the units of analysis. There is no reason to believe that these factor structures would be similar (i.e., that work styles would covary similarly across occupations as they do across), and assuming they do would be a variant on the well-known ecological fallacy (Robinson, 1950). Granted, in some domains, there is evidence of similarity when treating individuals versus occupations as the units of analysis (e.g., the emergence of RIASEC occupation interest themes at both the individual and occupational levels of analysis; Holland, 1997). However, for Work Styles, that is clearly not the case.

Table 6.8. Comparison of PCA-Based Higher-Order Work Style Dimensions and Higher-Order Work Style Dimensions from Phase 1

PCA-Based Higher Order Dimensions / Work Styles	Description	Higher-Order Work Style Dimensions from Phase 1
Proactive and Growth Oriented	A tendency to proactively pursue excellence through setting high standards, engaging in continuous learning, and persisting and adapting in the face of challenges and uncertainty at work.	
Innovation	A tendency to be inventive, to be imaginative, and to adopt new perspectives on ways to accomplish work.	Compound Dimension
Achievement Orientation	A tendency to establish and maintain personally challenging work-related goals, set high work-related standards, and exert high effort toward meeting those goals and standards.	Conscientiousness
Intellectual Curiosity	A tendency to seek out and acquire new work-related knowledge and obtain a deep understanding of work-related subjects.	Openness
Tolerance for Ambiguity	A tendency to be comfortable with ambiguity and uncertainty at work.	Openness
Initiative	A tendency to be proactive and take on extra responsibilities and tasks that may fall outside of one's required work role.	Compound Dimensions
Adaptability	A tendency to be open to and comfortable with change, new experiences, or ideas at work.	Openness
Self-Confidence	A tendency to believe in one's work-related capabilities and ability to control one's work-related outcomes.	Conscientiousness
Perseverance	A tendency to exhibit determination and resolve to perform or complete tasks in the face of difficult circumstances or obstacles at work.	Compound Dimensions
Leadership Orientation	A tendency to lead, take charge, offer opinions, and provide direction at work.	Extraversion
Interpersonally Oriented	A tendency to engage with others in a positive, supportive, selfless, and considerate manner at work.	
Humility	A tendency to be modest and humble when interacting with others at work.	Honesty-Humility
Sincerity	A tendency to be genuine and sincere in interactions with others at work, without concern for personal gain or self-interest.	Honesty-Humility
Empathy	A tendency to show concern for others and be sensitive to others' needs and feelings at work.	Agreeableness
Cooperation	A tendency to be pleasant, helpful, and willing to assist others at work.	Agreeableness
Optimism	A tendency to exhibit a positive attitude and positive emotions at work, even under difficult circumstances.	Compound Dimensions
Social Orientation	A tendency to seek out, enjoy, and be energized by social interaction at work.	Extraversion

Table 6.8. (Continued)

PCA-Based Higher Order Dimensions / Work Styles	Description	Higher-Order Work Style Dimensions from Phase 1
Conscientious and Rule Oriented	A tendency to approach work with careful attention to quality, reliability, and ethical standards while maintaining organized and thorough work practices.	
Cautiousness	A tendency to be careful, deliberate, and risk-avoidant when making work-related decisions or doing work.	Conscientiousness
Attention to Detail	A tendency to be detail-oriented, organized, and thorough in completing work.	Conscientiousness
Dependability	A tendency to be reliable, responsible, and consistent in meeting work-related obligations.	Conscientiousness
Integrity	A tendency to be honest and ethical at work.	Honesty-Humility
Emotionally Resilient	A tendency to maintain emotional stability and cope effectively in the face of adversity, criticism, or high-pressure situations at work.	
Stress Tolerance	A tendency to cope and function effectively in stressful situations at work.	Emotional Stability
Self-Control	A tendency to remain calm and composed and to manage emotions effectively in response to criticism or difficult situations at work.	Emotional Stability

Note. Higher-order dimensions are highlighted in light grey.

To further evaluate the substantive meaningfulness of these components, we produced component scores (using the standard, regression-based method) and summarized them by O*NET occupation job zone and job family (see Tables 6.9 and 6.10, respectively). Additionally, Table 6.11 lists the top five and bottom five occupations on each Work Style component.

The pattern of results in Tables 6.9 through 6.11 appears to lend further credence to the meaningfulness of basing higher-order Work Style dimensions on these components. For example, as one might expect, the Proactive and Growth Oriented component correlated strongly and positively with job zone, where job zone five occupations exhibited the highest Proactive and Growth Oriented component scores on average. This finding is consistent with the notion that the higher the job zone, the more extensive the education and preparation needed to perform in the occupation. As another example, examination of component scores by job family revealed that job families with the highest Interpersonally Oriented component scores tended to have clear interpersonal elements to them (e.g., Community and Social Service, Healthcare Support, Healthcare Practitioners and Technical, and Personal Care and Service). Job families with the highest Conscientiousness and Rule Oriented component scores (Legal) and highest Emotionally Resilient component scores (Protective Service) tended to be very rule-oriented and stress-laden, respectively. This pattern of findings is further reinforced by examining occupations with the highest and lowest scores on each component in Table 6.11 (e.g., Clergy and Mental Health Counselors as the top Interpersonally Oriented occupations, occupations involving working with nuclear technology as the top Conscientious and Rule Oriented occupations, Air Traffic Controllers and protective service occupations as the top Emotionally Resilient occupations).

Table 6.9. Mean Work Style Component Scores by Job Zone

Job Zone	*n*	Component			
		Proactive and Growth Oriented	Interpersonally Oriented	Conscientious and Rule Oriented	Emotionally Resilient
1	31	-1.29	-.90	-1.00	-.64
2	289	-.97	-.45	-.26	-.14
3	211	-.05	.05	.27	.11
4	211	.86	.11	.06	.15
5	149	.99	.84	.24	.06

Note. Component scores are on a z-score metric. Component scores for the Interpersonally Oriented and Emotionally Resilient scores were multiplied by -1, so that higher scores reflect greater impact of Interpersonal Orientation and Emotional Resilience on job performance, respectively.

Based on the observation and results above, the Center and project team agreed to use the PCA-based higher-order structure as the basis for updating the higher-order Work Styles dimensions in O*NET. The rationale for this was threefold: First, given that O*NET Work Style ratings are meant to be indicators of personality-related requirements of occupations, the team felt it was important that the higher-order structure reflects this intent as well. Second, the PCA-based higher-order dimensions provide a more accurate representation of the higher-order structure underlying the full set of active, data-level occupations, relative to the provisional structure identified in Putka et al. (2024). Lastly, the PCA-based higher-order dimensions exhibited substantively intuitive and meaningful differences across job zones and job families.

Table 6.10. Mean Work Style Component Scores by Job Family

Job Family	*n*	Component			
		Proactive and Growth Oriented	Interpersonally Oriented	Conscientious and Rule Oriented	Emotionally Resilient
Architecture and Engineering	53	.77	-.48	.74	-.57
Arts, Design, Entertainment, Sports, and Media	40	.83	-.35	-1.57	-.14
Building and Grounds Cleaning and Maintenance	8	-.76	-.45	-.68	-.23
Business and Financial Operations	45	.57	-.02	.40	.29
Community and Social Service	14	.54	2.15	-.45	.70
Computer and Mathematical	36	1.02	-.40	.74	-.63
Construction and Extraction	57	-1.05	-1.02	-.26	-.11
Educational Instruction and Library	62	.83	1.13	-.66	-.07
Farming, Fishing, and Forestry	12	-.79	-.89	-.47	.06
Food Preparation and Serving Related	16	-.87	.38	-1.27	-.24
Healthcare Practitioners and Technical	89	.49	1.15	.82	.55
Healthcare Support	19	-.75	1.39	.17	.05
Installation, Maintenance, and Repair	49	-.44	-.71	.31	-.26
Legal	7	.27	.21	1.07	.66
Life, Physical, and Social Science	56	.88	.07	.52	-.64
Management	45	1.04	.39	-.28	1.11
Office and Administrative Support	51	-1.00	.16	.10	-.37
Personal Care and Service	31	-.54	1.00	-.81	-.01
Production	104	-1.07	-1.13	.07	-.78
Protective Service	26	.02	-.04	.31	1.94
Sales and Related	21	.42	-.06	-1.92	.75
Transportation and Material Moving	50	-.83	-.30	.03	.59

Note. Component scores are on a z-score metric. Component scores for the Interpersonally Oriented and Emotionally Resilient scores were multiplied by -1, so that higher scores reflect greater impact of Interpersonal Orientation and Emotional Resilience on job performance, respectively.

Table 6.11. Top 5 and Bottom 5 Occupations for Each Work Style Component

Proactive and Growth Oriented						
Top 5				*Bottom 5*		
11-1011.00	Chief Executives	2.08		53-7063.00	Machine Feeders and Offbearers	-1.86
27-2012.00	Producers and Directors	1.88		47-3014.00	Helpers--Painters, Paperhangers, Plasterers, and Stucco Masons	-1.89
15-1221.00	Computer and Information Research Scientists	1.78		35-9021.00	Dishwashers	-1.91
15-1255.01	Video Game Designers	1.76		41-2012.00	Gambling Change Persons and Booth Cashiers	-1.93
17-2199.09	Nanosystems Engineers	1.74		51-3022.00	Meat, Poultry, and Fish Cutters and Trimmers	-1.94
Interpersonally Oriented						
Top 5				*Bottom 5*		
21-2011.00	Clergy	2.81		33-9099.02	Retail Loss Prevention Specialists	-1.53
21-1014.00	Mental Health Counselors	2.47		51-9195.03	Stone Cutters and Carvers, Manufacturing	-1.53
21-1021.00	Child, Family, and School Social Workers	2.38		33-9031.00	Gambling Surveillance Officers and Gambling Investigators	-1.54
21-1023.00	Mental Health and Substance Abuse Social Workers	2.38		47-5051.00	Rock Splitters, Quarry	-1.77
25-2051.00	Special Education Teachers, Preschool	2.34		45-4021.00	Fallers	-1.78
Conscientious and Rule Oriented						
Top 5				*Bottom 5*		
51-8011.00	Nuclear Power Reactor Operators	2.15		27-2032.00	Choreographers	-3.43
19-4051.00	Nuclear Technicians	2.12		41-9041.00	Telemarketers	-3.59
19-4051.02	Nuclear Monitoring Technicians	2.03		41-9012.00	Models	-4.00
17-2161.00	Nuclear Engineers	1.90		41-9091.00	Door-to-Door Sales Workers, News and Street Vendors, and Related Workers	-4.19
29-1222.00	Physicians, Pathologists	1.89		27-2011.00	Actors	-4.51
Emotionally Resilient						
Top 5				*Bottom 5*		
53-2021.00	Air Traffic Controllers	3.17		15-2021.00	Mathematicians	-1.83
33-3012.00	Correctional Officers and Jailers	2.96		27-1024.00	Graphic Designers	-1.84
33-1011.00	First-Line Supervisors of Correctional Officers	2.94		27-1013.00	Fine Artists, Including Painters, Sculptors, and Illustrators	-2.35
33-3052.00	Transit and Railroad Police	2.90		27-1012.00	Craft Artists	-2.54
33-1021.00	First-Line Supervisors of Firefighting and Prevention Workers	2.82		27-3043.05	Poets, Lyricists and Creative Writers	-2.66

Note. Each block of occupations includes the following: O*NET-SOC 2019 code, O*NET-SOC title, and score on the given component. Component scores are on a z-score metric. Component scores for the Interpersonally Oriented and Emotionally Resilient scores were multiplied by -1, so that higher scores reflect greater impact of Interpersonal Orientation and Emotional Resilience on job performance, respectively.

Identifying Distinctive Work Styles for Each Occupation

As part of Step 6, we also explored additional metrics to include alongside the Work Style impact ratings for each active, data-level occupation in the O*NET database. Specifically, the Center expressed interest in identifying Work Styles that helped distinguish each occupation. Such distinguishing, or "distinctive" Work Styles would be akin to RIASEC vocational interest "high point codes" for each occupation. However, unlike the RIASEC high-point codes, the objective here was not simply to identify the top-rated Work Styles for each occupation, but rather those Work Styles that met some minimum level of being beneficial to work in an occupation *and* that helped distinguish the occupations from others. The logic behind this objective was as follows: unlike the vocational interest domain, where each RIASEC dimension varies greatly in how characteristic it is of work across occupations, some Work Styles are nearly universally beneficial regardless of occupation (e.g., Dependability and Attention to Detail; see Table 6.1). Were the Center to adopt a strategy where it simply highlighted the top three rated Work Styles for each occupation, Dependability and Attention to Detail would dominate the "top 3" Work Style lists across occupations. As such, the Center sought to identify Work Styles that were clearly beneficial to an occupation but also helped distinguish that occupation from others, with the intent of aiding career exploration use cases of Work Style data.

In light of the need above, we developed a three-step process for identifying a subset of beneficial yet distinctive Work Styles to highlight as most distinctive for each occupation. The process is as follows:

- **Step 1**: Identify all Work Styles for an occupation that have ratings of at least 2.0 on the impact rating scale (i.e., rated at least as "beneficial" based on rating scale anchors).
- **Step 2:** Sort the Work Styles from Step 1 for an occupation in ascending order (lowest to highest) of the number of times the Work Style had an impact rating of 2.0 or above across all 891 active, data-level occupations. In the case of ties in these counts, sort the tied Work Styles in descending order (highest to lowest) of their impact rating for the given occupation.[14]
- **Step 3:** For each occupation, assign up to 10 of the Work Styles from Step 2 a "distinctiveness rank" based on their position in the list: the Work Style sorted first is rank 1, the next is rank 2, etc. Any Work Styles beyond position 10 are not ranked.

A "Distinctiveness Rank" field containing ranks for each data-level occupation will appear in the O*NET Work Styles data table starting with O*NET 30.1. Note, although the database provides up to 10 distinctive Work Styles for each occupation, this is primarily meant to give the Center and other users of O*NET flexibility in terms of how these data are used/rendered (e.g., for example for some use cases, one may only want to present the top 3 or top 5 most distinctive Work Styles for an occupation, which can readily be identified with these data).

[14] Note, it is possible that fewer than 10 Work Styles will emerge as beneficial for an occupation based on Step 1. In such cases, an occupation would have fewer than 10 distinctive Work Styles.

Summary of Distinctiveness Ranks

As noted above, given the way distinctiveness ranks are defined, an occupation can have fewer than 10 Work Styles with distinctiveness ranks, as only Work Styles with impact ratings of 2.0 or above for an occupation are eligible to have a distinctiveness rank for that occupation. Table 6.12 examines the prevalence of this issue. Specifically, it provides a count of the number of occupations that have 10 ranked Work Styles, at least nine ranked Work Styles, at least eight ranked Work Styles, etc. As Table 6.12 reveals, only 11.1% of occupations have 10 Work Styles with distinctiveness ranks. 46.0% of occupations have at least five Work Styles with distinctiveness ranks, and 72.0% of occupations have at least three Work Styles with distinctiveness ranks. At the extreme end, all but eight occupations have at least one Work Style with a distinctiveness rank (as defined in the process above).[15] This information can be of value to users of these data who are looking to present distinctiveness rank data, as it gives a sense of how dense (or sparse) the rank data are across the population of data-level occupations, depending on how far down the rank list the user may want to go. For example, if a user wants to ensure that a complete set of ranks appears for at least 70% of all occupations, then they should focus only on presenting the top three Work Styles by distinctiveness rank.

Table 6.12. Count of Occupations that have at least X Work Styles with Distinctiveness Ranks

X	*n*	%
10	99	11.1
9	139	15.6
8	191	21.4
7	263	29.5
6	335	37.6
5	410	46.0
4	515	57.8
3	645	72.4
2	819	91.9
1*	883	99.1

Note. % is out of 891 occupations. *The *n* and % displayed in this row do not include the eight occupations that initially did not have at least one Work Style with a distinctiveness rank.

To help illustrate the differentiating value the distinctiveness ranks have relative to impact ratings, Table 6.13 provides the count occupations where each Work Style was in the top ten most distinctive Work Styles for each occupation based on its distinctive rank, and for comparison, Table 6.14 provides the count occupations where each Work Style was in the top ten Work Styles based on impact ratings.

Focusing on the percentage of occupations (out of 891) for which each Work Style was in the top three Work Styles (see % Top 3 column in Tables 6.13 and 6.14), note that the distribution of % Top 3 values is much flatter when ranking is based on distinctiveness rather than impact

[15] For purposes of populating the Work Styles data in the O*NET database for these eight occupations, the Work Style with the highest impact rating for the occupation was assigned a distinctiveness rank of one. Doing so ensured that all active, data-level occupations had at least one Work Style with a distinctiveness rank. The impact ratings for the top Work Styles for these eight occupations were still relatively high even though they did not meet the 2.0 threshold (Range: 1.81 to 1.98).

rating. Indeed, comparing standard deviation, skew, and kurtosis values for the distribution of % of Top 3 values across Work Styles clearly indicated the distinctiveness ranks (Top 3% *SD* = 9.98, *Skew* = 1.81, *Kurtosis* = 3.24) more evenly spread out Work Styles relative to ranking Work Styles based on impact ratings (Top 3% *SD* = 23.99, *Skew* = 2.41, *Kurtosis* = 5.27).

Table 6.13. Count of Occupations Where Each Work Style was Ranked 1st through 10th based on Distinctiveness Rank

Work Style	% Top 3	Count of Occupations Where Work Style was Ranked...										Not Ranked
		1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
Attention To Detail	38.7	149	106	90	58	60	54	34	29	21	23	267
Dependability	38.5	53	169	121	100	70	72	72	50	40	29	115
Cautiousness	24.5	138	40	40	49	26	20	18	12	7	5	536
Achievement Orientation	14.1	18	65	43	24	22	21	14	18	4	3	659
Innovation	14.1	105	7	14	4	1	1	0	0	0	0	759
Intellectual Curiosity	13.1	24	31	62	31	27	17	10	13	4	13	659
Perseverance	11.9	32	39	35	19	7	0	0	0	0	0	759
Integrity	11.8	12	57	36	45	56	46	40	30	30	20	519
Leadership Orientation	10.9	46	26	25	7	0	0	0	0	0	0	787
Social Orientation	10.4	32	34	27	20	34	17	28	13	3	2	681
Stress Tolerance	10.3	30	30	32	19	30	19	5	2	0	0	724
Empathy	9.4	20	42	22	28	20	3	1	0	0	0	755
Optimism	9.2	54	24	4	0	0	0	0	0	0	0	809
Sincerity	9.2	47	34	1	0	0	0	0	0	0	0	809
Adaptability	8.9	10	35	34	35	14	5	1	0	0	0	757
Cooperation	8.0	18	27	26	39	22	27	23	20	28	4	657
Self-Confidence	7.3	38	24	3	0	0	0	0	0	0	0	826
Self-Control	7.3	9	26	30	37	21	33	17	4	2	0	712
Initiative	4.6	38	3	0	0	0	0	0	0	0	0	850
Tolerance For Ambiguity	0.9	8	0	0	0	0	0	0	0	0	0	883
Humility	0.2	2	0	0	0	0	0	0	0	0	0	889

Note. Work Styles are sorted in descending order of the percentage of occupations (out of 891) for which that Work Style was in the top three Work Styles based on distinctiveness ranks. The counts in this table do not include the eight occupations that initially did not have at least one Work Style with a distinctiveness rank.

Table 6.14. Count of Occupations Where Each Work Style was Ranked 1st through 10th based on Impact Rating

Work Style	% Top 3	Count of Occupations Where Work Style was Ranked...										Not Ranked
		1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
Dependability	88.9	267	368	157	52	16	14	7	4	1	2	3
Attention To Detail	72.4	372	173	100	48	18	22	25	22	19	19	73
Cautiousness	39.2	33	61	255	104	57	39	33	28	22	14	245
Integrity	19.1	32	46	92	157	120	130	116	58	32	29	79
Cooperation	16.8	18	58	74	61	56	57	61	66	58	71	311
Intellectual Curiosity	12.1	62	19	27	54	49	34	22	24	31	34	535
Innovation	9.8	20	31	36	21	22	15	24	15	14	22	671
Social Orientation	9.7	24	29	33	40	55	29	35	15	23	24	584
Empathy	6.7	41	5	14	17	31	29	18	23	18	12	683
Leadership Orientation	6.7	25	26	9	12	13	6	8	11	11	15	755
Stress Tolerance	4.5	18	5	17	87	83	91	89	66	64	71	300
Perseverance	3.3	6	8	15	78	96	76	110	111	87	76	228
Self-Control	2.6	8	4	11	27	68	105	91	117	86	63	311
Achievement Orientation	2.4	2	4	15	72	119	124	114	77	69	54	241
Optimism	2.2	3	7	10	25	21	20	20	19	25	20	721
Adaptability	1.3	1	2	9	21	30	40	44	78	105	89	472
Self-Confidence	1.2	2	3	6	7	19	20	22	81	105	119	507
Sincerity	0.9	0	3	5	7	8	19	28	26	45	41	709
Tolerance For Ambiguity	0.2	0	1	1	0	3	6	11	11	18	29	811
Initiative	0.1	0	0	1	0	5	13	13	34	55	78	692
Humility	0.0	0	0	0	0	2	2	0	5	3	9	870

Note. Work Styles are sorted in descending order of the percentage of occupations (out of 891) for which that Work Style was in the top three Work Styles based on impact ratings. The counts in this table do not include the eight occupations that initially did not have at least one Work Style with a distinctiveness rank.

Step 7: Finalizing Work Style Ratings O*NET 30.1

The Work Style ratings produced as part of Step 6 for all 891 active, data-level occupations in O*NET 29.3 could be considered draft final ratings for the O*NET 30.1 database. To finalize those ratings for inclusion in O*NET 30.1, we constructed a file comprising occupational data from O*NET 29.3, draft final LLM ratings for each Work Style-occupation combination, and a set of diagnostic flags to highlight draft final LLM ratings for Work Style-occupation combinations that require closer inspection. The purpose of constructing this file was to facilitate a final review of the draft final Work Style ratings by our primary expert consultant on this work, Dr. Fred Oswald.

In terms of occupational data from O*NET 29.3, the following data were provided in the final review file:

- Occupation O*NET-SOC code
- Occupation title
- Occupation description
- Occupation task statements (core when available, otherwise all)
- Occupation sample of reported job titles

In addition to the information above from O*NET 29.3, the following data were included in the final review file:

- Draft final Work Style-occupation impact ratings (based on using O*NET 29.3 occupation data as input).
- Expert Work Style ratings for the subset of 125 occupations from Step 3.
- A set of diagnostic flags indicating Work Style-occupation combinations flagged for closer review (detailed below).
- Z-scores of O*NET 29.3 Work Style ratings for those Work Styles that have direct analogues to those in the new set of Work Styles.
- Within-LLM z-scores for each LLM composite rating (i.e., z-score for the composite of two Claude-based ratings and one Llama-based rating). In this case, *z*-scores were calculated across the 891 data-level occupations in 29.3.

We provided Dr. Oswald with the review file above that contained draft final impact ratings for all 21 Work Styles on the 891 occupations above (18,711 ratings in total). Of these 18,711 ratings, we flagged 575 (3.1%) for closer review based on three sets of flagging rules described below:

- First, among the 2,499 Work Style-occupation combinations where we had both draft final LLM and expert ratings, we flagged Work Style-occupation combinations where the final LLM rating differed from the expert rating by one or more points on the 7-point (-3 to +3 rating scale). We flagged 43 of 18,711 Work Style-occupation combinations (.2%, 1.7% of combinations for which we had both sources of ratings) for this reason.

- Second, we flagged Work Style-occupation combinations that had highly different ratings among the three average LLM ratings that comprised the draft final rating. Specifically, we flagged a Work Style-occupation combination if the *z*-scores for the LLM composites that comprised the Work Style-occupation combinations' draft final rating included both a negative and a positive z-score with a difference of 1.5 or greater. We flagged 277 of 18,711 (1.5%) of Work Style-occupation combinations for this reason. We also flagged a Work Style-occupation combination if the *z*-scores for the LLM composites that comprised the occupation-Work Style combinations had a range of 2.0 or greater (regardless of the sign of those *z*-scores). We flagged 147 of 18,711 (.8%) of Work Style-occupation combinations for this reason.

- Lastly, for those Work Styles that had an analog to Work Styles with data in the published O*NET 29.3 database, we flagged Work Style-occupation combinations whose draft final LLM rating *z*-score differed from the published Work Style rating *z*-score by 2.0 or more. We flagged 310 of 18,711 (1.7%) occupation-Work Style combinations for this reason.

The thresholds above were determined by visual inspection of the distribution of the variables involved in setting the thresholds above (e.g., LLM-expert differences, differences among LLM composite z-scores, differences among draft LLM z-scores, and published Work Style *z*-scores).

Dr. Oswald was tasked with reviewing the full set of ratings, with particular attention paid to the flagged Work Style-occupation ratings, and making updated final ratings where he believed the draft final LLM rating needed to be adjusted based on his review. Specifically, we offered the following general suggestions for review:

- Consider one Work Style at a time and first examine the occupations flagged for review for that Work Style and why the given occupation was flagged.

- Next, consider viewing where the flagged occupation for a given Work Style falls relative to all other occupations for that Work Style. Ask yourself, does the draft LLM rating for that flagged occupation seem higher or lower than it should be relative to other occupations adjacent to it in the list? Consult the definition of the Work Style and the occupation description of tasks when doing so (also provided as part of the review materials).

- Next, consider viewing where the flagged occupation for a given Work Style falls relative to other occupations *within its job family*. Ask yourself, does the draft LLM rating for that flagged occupation seem higher or lower than it should be relative to other occupations in its job family? Again, consult the definition of the Work Style and the occupation description of tasks when doing so (also provided as part of the review materials).

- Ultimately, view your role here as serving as the “final” expert reviewer/gatekeeper for Work Style ratings that will be published to the O*NET database for all data-level occupations. So consider the suggestions above as a guide, but not necessarily exhaustive of everything you may want to consider as you review the ratings.

Results of Review

Of the 18,711 draft final LLM ratings, 249 of them were adjusted by Dr. Oswald (1.3%), and all of those were from the set of 575 Work Style-occupation combinations flagged for review. In general, the magnitude of adjustments was very small on average (*M* = -01 [final-draft], *SD* = .34). Table 7.1 provides a frequency distribution of the differences between final (adjusted) LLM ratings and the draft final ratings produced by Step 6. As shown in Table 7.1, 90% of disagreements fell in the range of -.55 to .44, indicating that when LLM ratings were adjusted based on expert review, they were largely within .5 of a scale point of the draft rating.

Table 7.1. Summary of Differences between Final and Draft Work Style Ratings

Statistic	Final Rating - Draft Rating
Mean	-.01
SD	.34
Min	-1.58
Max	2.43
Percentiles	
5	-.55
25	-.10
50	-.01
75	.07
95	.99

Note. n = 249 (occupation-Work Style combinations where the difference between final and draft ratings was not zero). Positive values indicate final ratings were higher than draft ratings, and negative values indicate that draft ratings were higher than final ratings.

When asked to summarize his approach and general thoughts on the review, Dr. Oswald indicated that his strategy involved first filtering by each Work Style and then, within a given Work Style, filtering by each job family and ranking occupations within the family in ascending order according to the draft final LLM rating. This was to allow for Work Style-specific comparisons. There were a couple of cases where he expanded his consideration beyond the focal Work Style: (a) when there were not many jobs (this was not frequent, but did happen), (b) when the Work Style-occupation flags were at the extreme ends of the draft final LLM rating continuum (so that he could anchor those flags at these endpoints against the wider Work Style rating continuum across jobs). Next, he indicated that he would investigate the reason why the occupation was flagged for the given Work Style (e.g., expert-LLM disagreement, disagreement among LLM composites), yet he did not engage in vote-counting or a similar mechanical strategy to resolve the given discrepancy. However, he did note three observations he made during this process that speak to the quality of the draft final LLM ratings. First, when there was an LLM-expert disagreement, it was *almost always the LLM that was (in his opinion) correct*. Second, there were exceptions, when the expert rating was correct (or partially so, where the best rating seemed to lie between LLM and expert); this was rare but not to be dismissed. Lastly, even more rare, were cases where his final rating for a Work Style departed from both expert and draft final LLM rating for an occupation – this was sometimes because the occupation in question was similar to other occupations (similar through the lens of how the Work Style manifests in the occupations, or similar in terms of full job description) and that similar occupation fell somewhere else on the continuum of ratings for that Work Style.

Beyond an overall summary, we also examined differences by Work Style and job family to see if any notable trends emerged. Tables 7.2 and 7.3 present the results of these analyses, again limiting analyses to only those 249 draft ratings that were adjusted during final review. As shown in these tables, differences appeared small and varied little by Work Style or job family. We strongly urge caution over interpreting any observed differences in Tables 7.2 and 7.3 due to the very small sample sizes on which the descriptive statistics for each Work Style and job family are based.

Table 7.2. Summary of Differences between Final and Draft Work Style Ratings by Work Style

Work Style		Final Rating - Draft Rating			
	n	*M*	*SD*	*Min*	*Max*
Achievement Orientation	12	-.23	.54	-1.58	.31
Adaptability	20	-.22	.44	-1.30	.58
Attention To Detail	19	.08	.28	-.33	.88
Cautiousness	7	.25	.98	-.50	2.43
Cooperation	12	-.05	.39	-.60	.77
Dependability	37	.00	.07	-.15	.24
Empathy	3	-.03	.01	-.04	-.02
Humility	24	.00	.19	-.59	.45
Initiative	8	-.02	.13	-.24	.18
Innovation	7	-.03	.37	-.68	.39
Integrity	13	.16	.50	-.58	.89
Intellectual Curiosity	0	-	-	-	-
Leadership Orientation	3	-.02	.04	-.05	.02
Optimism	7	-.09	.09	-.22	.03
Perseverance	17	.07	.10	-.10	.22
Self-Confidence	4	.05	.12	-.03	.23
Self-Control	9	.03	.14	-.14	.29
Sincerity	12	-.09	.07	-.20	.06
Social Orientation	8	.18	.62	-.11	1.70
Stress Tolerance	10	-.03	.31	-.85	.34
Tolerance For Ambiguity	17	.01	.27	-.59	.88

Note. n = Number of occupations where the difference between final and draft ratings for the given Work Style was not zero. Positive values indicate final ratings were higher than draft ratings, and negative values indicate that draft ratings were higher than final ratings.

Table 7.3. Summary of Differences between Final and Draft Work Style Ratings by Job Family

Job Family	*n*	Final Rating - Draft Rating			
		M	*SD*	*Min*	*Max*
Architecture and Engineering	8	-.14	.68	-1.58	.88
Arts, Design, Entertainment, Sports, and Media	24	.22	.33	-.19	.89
Building and Grounds Cleaning and Maintenance	1	-.22	-	-.22	-.22
Business and Financial Operations	11	-.05	.16	-.50	.14
Community and Social Service	1	-.34	-	-.34	-.34
Computer and Mathematical	3	.10	.15	-.06	.24
Construction and Extraction	25	-.02	.15	-.52	.23
Educational Instruction and Library	4	.06	.11	-.04	.21
Farming, Fishing, and Forestry	4	-.23	.50	-.85	.34
Food Preparation and Serving Related	3	-.07	.46	-.60	.22
Healthcare Practitioners and Technical	11	.11	.80	-.68	2.43
Healthcare Support	3	-.13	.13	-.27	-.02
Installation, Maintenance, and Repair	14	-.05	.15	-.33	.20
Legal	4	-.19	.27	-.59	.01
Life, Physical, and Social Science	12	-.19	.40	-1.15	.11
Management	8	.02	.28	-.41	.58
Office and Administrative Support	32	.02	.29	-1.30	.43
Personal Care and Service	13	-.05	.12	-.24	.10
Production	29	-.01	.16	-.59	.37
Protective Service	17	.12	.54	-.77	1.70
Sales and Related	11	-.06	.09	-.20	.09
Transportation and Material Moving	11	-.16	.22	-.58	.07

Note. n = Number of occupation-Work Style combinations within the given job family where the difference between final and draft ratings was not zero. Positive values indicate final ratings were higher than draft ratings, and negative values indicate that draft ratings were higher than final ratings.

Summary

Based on the results of Dr. Oswald's review, a few changes were made to the draft final set of LLM ratings for purposes of finalizing them for inclusion in O*NET 30.1. These changes were implemented, and the distinctiveness ranking process described in Step 6 was re-run based on the updated ratings to arrive at a final set of Work Style impact ratings and distinctiveness rank data for the 891 active, data-level O*NET-SOCs. The final set of impact ratings and distinctiveness ranks produced here will be published in O*NET 30.1 database, targeted for release in December 2025.

Guidance for Updating Work Style Ratings in Future Versions of the O*NET Database

To facilitate the application of the hybrid generative AI-expert Work Style rating process developed here in future versions of the O*NET database, we suggest the following process:

1. Download the following data from the latest available version of the O*NET database:
    - Occupation O*NET-SOC code
    - Occupation title
    - Occupation description
    - Occupation task statements and emerging tasks (to identify “New” and “AI/SME” emerging tasks)
    - Occupation sample of reported job titles
    - Occupation Work Style impact ratings and distinctiveness ranks
2. Download all of the data above (sans Work Style data) from the O*NET 29.3 archival data release and have available the expert ratings of the 21 Work Styles evaluated as part of the current project that were based on O*NET 29.3 data.
3. Confirm the three LLMs used to generate LLM-based ratings in the current effort are still available through the AWS Bedrock service (assuming that service is used in future updating efforts). If they are not, engage in the following process to evaluate the implications of dropping/or replacing the LLMs that are no longer available:
    a. Identify one or more best-bet replacement LLMs for evaluation that are available via AWS Bedrock.
    b. Prompt the new model three times using the final prompt template provided in Appendix B and evaluate convergence between average rating produced by that model (across three runs) and expert data analyzed in the current study based on O*NET 29.3 (i.e., use O*NET 29.3 for the 125 occupations evaluated in Step 5 in this study and input to the prompt to maintain comparability to expert ratings).
    c. Update the set of LLMs used to generate ratings to reflect inclusion of new LLMs added (and removal of LLMs dropped) to include updating norms still using expert ratings on O*NET 29.3 data as a reference point (via the process described in Appendix C).
4. Apply the updated rating approach to all data-level occupations in the latest available O*NET data (downloaded in #1) to generate a draft updated set of Work Style impact ratings and distinctiveness ranks that are based on the latest available O*NET data.
5. Merge the updated version of the Work Style impact ratings (from #4) and the Work Style ratings from the latest available published Work Style ratings, and create the following occupation-level flags to identify those occupations where a closer review by an external personality-job analysis expert is most critical:

a. Work Style-occupation combinations where the updated impact ratings and the latest available impact ratings differ by one or more points on the 7-point impact rating scale (-3 to +3 rating scale).

b. Work Style-occupation combinations that had highly different ratings among the LLM rating composites that underly the new impact ratings. Specifically, Work Style-occupation combinations where *z*-scores for the LLM composites that comprised updated ratings met either of the following conditions: (a) the updated set of LLM rating composites included a negative and positive *z*-score with a difference between them of 1.5 or greater, or (b) the z-scores for the LLM composites that comprised the updated ratings had a range of 2.0 or greater (regardless of sign of those *z*-scores).

Note that the flags above directly correspond to the first two sets of criteria used to identify Work Style-occupation combinations for expert review in Step 7.

c. All Work Style ratings for occupations that did not have Work Style impact ratings in the latest available version of the O*NET database (i.e., new occupations).

6. Have an external expert review the updated set of Work Style-occupation ratings with particular attention being paid to Work-Style-occupation combinations flagged for review above. The flags above are intended to draw an expert's attention to those Work Style-occupation combinations where the updated impact ratings are least aligned with the latest available ratings. The personality-job analysis expert tasked with providing a final review of updated impact ratings before they go "live" to the operational O*NET database would adjust the impact ratings as needed based upon their review (similar to the final review conducted as part of Step 7 in the current effort).

7. Following Work Style expert review and revision of the impact ratings, update the distinctiveness ranks (a direct function of the impact ratings for each occupation) and provide the final updated Work Style table to the Center for final review prior to being incorporated into the operational O*NET Database.

Timing of Future Work Style Rating Updates

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

- The Center may consider making updates to Work Style ratings *every time* information for an occupation that serves as input to the LLM-based method changes (i.e., occupation title, description, tasks, sample of reported job titles), or a new occupation is introduced into the O*NET Database. The update process above could be carried out following the introduction of these new inputs or occupations, and Work Style ratings for the relevant occupations would be incorporated into a subsequent version of the O*NET Database.

- The Center may consider making updates to the interest data database-wide every *X* years (e.g., every 1 or 2 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

Work Style ratings 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 LLM-based prediction method (i.e., occupation title, description, tasks, sample of reported job titles) change for existing occupations. When discussing potential updating options for existing occupations with the Center, it was noted that most changes would likely be to samples of reported job titles, as mentioned in the inputs above. 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 of new task statements being added to listings between Work Style ratings updates.

Lastly, when running updates for Work Styles in the future, the Center may also want to consider evaluating the latest available LLMs to see if they can improve upon the ratings generated by LLMs in the current effort. That is, beyond simply needing to consider new LLMs as a replacement for an current LLM in the event it is deprecated when future updates need to occur, it is possible that considering new LLMs may be able to further improve LLM-based ratings' convergence with expert ratings observed in the current report. Thus, there is also the potential for continuous improvement of ratings based on advances expected to occur with generative AI.

Conclusions and Future Directions

This report summarizes the successful effort to leverage advances in generative AI, combined with expert human judgment, to populate Work Style ratings for 891 data-level O*NET-SOC occupations. The generative AI methods used occupational information published within the O*NET database as input for generating Work Style ratings for occupations. As the world of work changes, these models can be applied to future versions of the O*NET database to quickly and efficiently generate and maintain high quality Work Style information for the O*NET system, which helps minimize the need for human raters.

For each of the 21 Work Styles, we developed an LLM prompt template that allowed us to generate Work Style ratings that converged well with ratings made by experts in personality and job analysis (i.e., highly experienced academic researchers who publish in this area) and ratings made by O*NET occupational analysts. The level of correlations we observed between LLM-based ratings and expert ratings approached levels of interrater reliability seen among expert raters. The LLM-based approach we developed eliminates the need to gather Work Style ratings via traditional expert or analyst data collections, helping to ensure updated, accurate Work Style information is available on a timely basis for O*NET customers and stakeholders.

Although our LLM-based approach performed well, we still advise the Center to incorporate a layer of expert review into LLM-based Work Style ratings. As part of this report, we've provided a recommended process for updating Work Style ratings over time that includes a step for expert review.

In terms of future directions, the success of the LLM-based approach here bodes well for potentially extending this work to automate the generation of ratings for other domains of

O*NET that have historically relied on high public burden or labor and time intensive forms of data collection with incumbents or occupational analysts (e.g., establishment-method-based work context data collections and occupational analyst-based abilities data collections). The methods examined here suggest an LLM-based approach could be a promising avenue to pursue for this purpose.

References

Anthropic. (2024). *The Claude 3 model family: Opus, sonnet, haiku.* Anthropic. Retrieved from https://www-cdn.anthropic.com/f2986af8d052f26236f6251da62d16172cfabd6e/claude-3-model-card.pdf

Borman, W. C., Kubisiak, U. C., & Schneider, R. J. (1999). Work styles. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 213–226). American Psychological Association.

Borman, W. C., McKee, A. S., & Schneider, R. J. (1995). Work styles. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *Development of prototype Occupational Information Network (O*NET) Content Model.* (Vol. 1, pp. 619-670). Utah Department of Workforce Services.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... Zhao, Z. (2024). *The Llama 3 Herd of Models.* arXiv. https://doi.org/10.48550/arXiv.2407.21783

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

Horn J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. https://doi.org/10.1007/BF02289447

Human Resources Research Organization (2025). Meet RRObot: HumRRO's secure, proprietary AI platform. https://www.humrro.org/corpsite/article/meet-rrobot-humrros-secure-proprietary-ai-platform/

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., ... El Sayed, W. (2023). *Mistral 7B.* arXiv. https://doi.org/10.48550/arXiv.2310.06825

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... El Sayed, W. (2024). *Mixtral of Experts.* arXiv. https://doi.org/10.48550/arXiv.2401.04088

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

National Center for O*NET Development (2025). *O*NET 30.0 Database.* O*NET Resource Center. Retrieved October 28, 2025, from https://www.onetcenter.org/database.html

Putka, D. J., Kell, H. J., Voss, N., Oswald, F., & Lewis, P. (2024). *Revisiting the Work Styles domain of the O*NET Content Model*. National Center for O*NET Development. https://www.onetcenter.org/reports/Work_Styles_New.html

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

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351-357. https://doi.org/10.2307/2087176

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Li, QV., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837. https://arxiv.org/pdf/2201.11903

Appendix A: Work Style Rating Materials for Experts and Analysts

Background

As you are aware, the Occupational Information Network (O*NET) is a comprehensive conceptual framework that provides a foundation for various human resource programs, including school curriculum development, job placement, and training. The National Center for O*NET Development has contracted with HumRRO to revise the Work Styles domain of the O*NET Content Model. Work Styles are heretofore defined as “personality tendencies exhibited at work that can affect how well someone performs a job.”

The current phase of research involves developing and evaluating potential methods for obtaining new Work Style ratings for O*NET occupations. These methods include Expert ratings (the gold standard), analyst ratings, LLM prompt-based ratings, and empirical ratings based on Work Style-GWA and Work Style-Work Context linkages. Your task, as expert raters, involves making work Style ratings for a subset of 125 O*NET occupations against which all other rating methods will be compared.

Your task, along with four of your colleagues, is to rate how beneficial (or detrimental) each of 21 Work Styles is to job performance in each of the 125 occupations.

Materials

1. Rating Instructions and Rating Scale (this document)
2. Rating Workbook (x2)
3. Stimulus Workbook (x2)

The two Rating Workbooks are used to capture your Work Style-occupation judgments at various points in the data collection effort. The first Rating Workbook, which includes 25 occupations, will be used during training and recalibration. The second Rating Workbook with 100 occupations will be used when you independently make your ratings after the virtual recalibration meeting. You were also provided two Stimulus Workbooks, which contain occupationally relevant information (i.e., descriptions, tasks, skills, Work Activity-Work Style linkages, Work Context-Work Style linkages) to allow you to make your ratings. The Stimulus Workbook with 25 occupations corresponds with the first 25 occupations in your first Rating Workbook, to be used during training and calibration. Similarly, the Stimulus Workbook, which contains 100 occupations, provides occupationally relevant information corresponding to the Rating Workbook, to be used post-recalibration (i.e., when independently rating the remaining subset of 100 occupations). Below, we provide additional guidance on how to use the materials provided when making your ratings.

Rating Scale

We have adapted a rating scale from the personality-oriented job analysis literature that you will use to rate how beneficial (or detrimental) each Work Style is to job performance in an occupation. When making your ratings, consider a person who has a relatively high standing in the given trait and assess whether that high standing is beneficial or detrimental to performance in the given occupation.

-3	-2	-1	0	+1	+2	+3
Very detrimental to job performance in this occupation	Detrimental to job performance in this occupation	Somewhat detrimental to job performance in this occupation	Little or no impact on job performance in this occupation	Somewhat beneficial to job performance in this occupation	Beneficial to job performance in this occupation	Very beneficial to job performance in this occupation
	WS: Humility--WC: Level of Competition (1.5)	WS: Cautiousness--WA: Thinking Creatively	WS: Empathy--WA: Analyzing Data or Information	WS: Humility--WA: Communicating with People Outside the Organization	WS: Sincerity--WA: Developing and Building Teams	WS: Leadership Orientation--WC: Coordinate or Lead Others
		WS: Empathy--WC: Level of Competition	WS: Humility--WA: Scheduling Work and Activities	WS: Self-Confidence--WA: Developing and Building Teams	WS: Initiative--WC: Unstructured Work	WS: Self-Control--WC: Conflict Situations
		WS: Sincerity--WC: Level of Competition	WS: Intellectual Curiosity--WA: Performing General Physical Activities	WS: Stress Tolerance--WA: Providing Consultation and Advice to Others	WS: Optimism--WA: Selling or Influencing Others	WS: Attention to Detail--WC: Importance of Being Exact or Accurate
		WS: Humility--WA: Selling or Influencing Others	WS: Sincerity--WA: Handling and Moving Objects	WS: Attention to Detail--WA: Selling or Influencing Others	WS: Attention to Detail--WA: Judging the Qualities of Objects, Services, or People	WS: Attention to Detail--WA: Inspecting Equipment, Structures, or Materials
		WS: Cooperation--WC: Level of Competition	WS: Leadership Orientation--WA: Controlling Machines and Processes	WS: Integrity--WC: Unstructured Work	WS: Social Orientation--WC: Public Speaking	WS: Empathy--WA: Assisting and Caring for Others
		WS: Humility--WC: Deal With Physically Aggressive People	WS: Social Orientation--WA: Repairing and Maintaining Electronic Equipment	WS: Optimism--WC: Face-to-Face Discussions	WS: Perseverance--WC: Time Pressure	WS: Innovation--WA: Thinking Creatively

Note. The values in the cells of the table above provide examples of work style - types of work (work activities, work contexts) pairs that fall at the given rating level.

Instructions

1. Start by carefully reviewing and familiarizing yourself with the definitions of Work Styles, Skills, Work Activities, and Work Contexts included in the Stimulus Workbook. We advise frequently returning to these definitions as you make your ratings.

2. Next, review the information for the occupation you are rating in the occupational description tab:
    a. Job zone will give you an idea of the level of preparation needed to enter the occupation.
    b. Focus on the occupation more generally and not with respect to performing in a specific occupational role or industry.
        i. For example, if Industrial-Organizational (I-O) Psychologists were one of the target occupations, do not think about different types of I-O positions (e.g., I-Os working in academe, tech, nonprofit, government, etc.), think of the I-O occupation as it would be described in the stimulus materials.

3. Review the task importance information for the occupation to get a sense for the tasks performed.

4. Review the skill importance information for the occupation
    a. Note, we're providing skill information because it helps complement task information, particularly with regard to the more interpersonal elements of the work in the occupation.

5. After completing steps 1-4, start to give thought to how you would rate the Work Styles for the occupation in question. This doesn't have to be anything formal. Rather, aim to have a mental sense of how you would rate the Work Styles based on the information above.

6. Once you give thought to how you may rate Work Styles for the occupation, review the WS-GWA and WS-WC linkages for the occupation to get a sense of how important general Work Activities and important Work Contexts for an occupation relate to Work Styles - recall these linkages were developed in an earlier phase of the Work Styles research. Use these to refine your initial ratings as warranted. Note, you do not need to change your initial ratings based on this information if you feel your initial ratings are accurate, but it may help you see things you didn't think of when initially rating Work Styles for an occupation.

7. Once you review the information described in #2 through #6 for an occupation, make your ratings for each of the 21 Work Styles for that occupation in the Ratings Workbook.

8. Repeat steps 2-6 until all occupations for a given assignment are rated. Once your assignment is complete, email a copy of your Rating Workbook to xx (xx@humrro.org) by the assigned completion date. Append the file name with your initials (e.g., Rating Workbook_XX.xlsx)

Tips for Completing Your Ratings and Important Points to Keep in Mind

1. To the extent the work in the occupation does NOT allow for the expression of the Work Style, you should lean towards rating it a zero.
2. To the extent the work in the occupation does allow for expression of the Work Style, it would open up the possibility to rate it as more beneficial or more detrimental.
3. To the extent there is more of a correspondence content-wise between the given Work Style and work in a given occupation, it would suggest more extreme ratings (i.e., trending towards either end of the scale).
4. Work Styles will inevitably be beneficial for several occupations, but you should try to differentiate across occupations to the extent possible
   - *Example*: Social Orientation may be a beneficial Work Style for several occupations, but for some, it would clearly be more beneficial (e.g., Sales)
   - *Example*: Self Control may be a beneficial Work Style for several occupations, but for some, it would clearly be more beneficial (e.g., Mediators)
5. Be aware of how your own personality and idiosyncratic experiences with an occupation may be influencing your ratings and aim to minimize their influence on your ratings – really try to ground your judgments in the way the occupation is described in the stimulus material provided.
6. To the extent you can, try to avoid allowing one-off counterfactuals to overly influence your ratings, such as: "I know someone who was low on Work Style X, but they were outstanding in Occupation Y, so I can't rate this Work Style as extremely beneficial." Instead, try to consider more typical cases and not focus on potential rare exceptions that may be more specific to your own experiences.
7. Recognize that Work Styles will likely vary in their relevance to tasks performed within an occupation. When this is the case, try not to overly focus on one single task or aspect of working in that occupation (e.g., a single work context or important skill); rather, let your ratings reflect a "mental averaging" that considers work in that occupation in the aggregate.
8. Remember the focus here is on whether a given Work Style is beneficial/detrimental for the performance in a given occupation. The focus is not on whether the Work Style would increase the likelihood of a person staying in or leaving an occupation, or being satisfied/dissatisfied with an occupation.
9. When thinking about "performance" in an occupation, our focus is more on task performance and the formally prescribed elements of the given occupation, and NOT on Organization Citizenship Behavior (OCB), contextual performance, and Counterproductive Work Behavior (CWB). Please keep this in mind as you make your ratings for each occupation.
10. Remember, review the WS-GWA and WS-WC linkages for an occupation only after you've reviewed the occupational description, tasks, and skill information for the occupation and

have given thought to how you would rate Work Styles for that occupation. Use the linkage information to refine your initial rating only as you see appropriate.

Appendix B: Final Prompt Template for Best-Bet LLM Rating Method

This appendix outlines the final prompt template and input variables for the best-bet LLM rating method developed in Step 3. It also includes a table of the custom few-shot content used for each Work Style.

Figure B.1. Final Prompt Template for the Best-Bet LLM Rating Method

You are an industrial-organizational psychologist who is an expert in personality measurement and job analysis. You will read a brief description of a personality attribute, a job description, and that job's most important tasks.

[Personality Attribute Name: Personality Attribute Description]: {WorkStyleName}: {WorkStyleDescription}

[Job Title: Job Description]:{JobTitle}:{JobDescription}{ExampleJobTitles}

[Job Tasks]:{JobTasks}

Your task is to rate the impact of the given personality attribute would have on an individual's performance on the given job using a 5 to 35 point scale, where rating points are defined as follows:

5: The personality attribute is very detrimental to performance of the job

10: The personality attribute is detrimental to performance of the job

15: The personality attribute is somewhat detrimental to performance of the job

20: The personality attribute has little or no impact on the performance of the job

25: The personality attribute is somewhat beneficial to performance of the job

30: The personality attribute is beneficial to performance of the job

35: The personality attribute is very beneficial to performance of the job

Any integer in the range 5 to 35 is allowed. Assign whatever rating you believe best reflects the degree to which this attribute hinders or helps performance in this job.

{CustomFewShot}

Please output your rating as a number that appears between XML tags <Rating> and </Rating>.

Note: This prompt template uses two types of formatting syntax: (1) Curly braces {} indicate variable placeholders that are dynamically replaced with specific content for the given Work Style-occupation combination to be rated (e.g., {WorkStyleName} is populated with the name of the Work Styles to be rated, such as " Achievement Orientation" or " Attention to Detail"); and (2) Angle brackets <> define XML tags that structure the LLMs output format, enabling programmatic extraction of the numeric rating for the given Work Style-occupation combination to be rated.

Table B.1. Variable Input Fields, References, and Example for the Final Prompt

Input Field	Description	O*NET Database Version	Example
WprkStyleName	Work Style Element Name		Achievement Orientation
WorkStyleDescription	Work Style Description		A tendency to establish and maintain personally challenging work-related goals, set high work-related standards, and exert high effort toward meeting those goals and standards.
JobTitle	O*NET Occupation Title	O*NET 29.3	Chief Executives
JobDescription	O*NET Occupation Description	O*NET 29.3	Determine and formulate policies, providing overall direction for companies or private and public sector organizations within guidelines established by a board of directors or a similar governing body. Plan, direct, or coordinate operational activities at the highest level of management, utilizing the assistance of subordinate executives and staff managers.
ExampleJobTitles	Sample of Reported Titles*	O*NET 29.3	Incumbents on this job sometimes have one of the following titles: CEO (Chief Executive Officer), Chief Financial Officer (CFO), Chief Operating Officer (COO), or President.
JobTasks	O*NET Occupation Task Statements**	O*NET 29.3	Direct or coordinate an organization's financial or budget activities to fund operations, maximize investments, or increase efficiency. Confer with board members, organization officials, or staff members to discuss issues, coordinate activities, or resolve problems. Prepare budgets for approval, including those for funding and implementing programs. Direct, plan, or implement policies, objectives, or activities of organizations or businesses to ensure continuing operations, to maximize returns on investments, or to increase productivity. Prepare or present reports concerning activities, expenses, budgets, government statutes or rulings, or other items affecting businesses or program services. Implement corrective action plans to solve organizational or departmental problems. Analyze operations to evaluate the performance of a company or its staff in meeting objectives or to determine areas of potential cost reduction, program improvement, or policy change. Direct or coordinate activities of businesses or departments concerned with the production, pricing, sales, or distribution of products. Direct human resources activities, including the approval of human resource plans or activities, the selection of directors or other high-level staff, or the establishment or organization of major departments. Appoint department heads or managers and assign or delegate responsibilities to them. Interpret

Input Field	Description	O*NET Database Version	Example
			and explain policies, rules, regulations, or laws to organizations, government or corporate officials, or individuals. Review reports submitted by staff members to recommend approval or to suggest changes. Negotiate or approve contracts or agreements with suppliers, distributors, federal or state agencies, or other organizational entities. Establish departmental responsibilities and coordinate functions among departments and sites to ensure seamless operations. Deliver speeches, write articles, or present information at meetings or conventions to promote services, exchange ideas, or accomplish objectives. Serve as liaisons between organizations, shareholders, and outside organizations. Coordinate the development or implementation of budgetary control systems, recordkeeping systems, or other administrative control processes. Preside over, or serve on, boards of directors, management committees, or other governing boards. Attend and participate in meetings of municipal councils or council committees. Organize or approve promotional campaigns.
CustomFewShot	Custom Few-Shot Example for each Work Style, consisting of O*NET Occupation Titles	See Table B.2.	When making your rating, consider the following examples: Examples of jobs where Achievement Orientation was rated 35 by experts: Climate Change Policy Analysts; Door-to-Door Salespeople; Fundraising Managers. Examples of jobs where Achievement Orientation was rated 32-33 by experts: Biochemists and Biophysicists; Kindergarten through Secondary Education Administrators; Postsecondary Chemistry Teachers; Financial Risk Specialists; Penetration Testers. Examples of jobs where Achievement Orientation was rated 30 by experts: Computer Systems Analysts; Cost Estimators; Geodetic Surveyors; Soil and Plant Scientists; Water Resource Specialists. Examples of jobs where Achievement Orientation was rated 27-28 by experts: Registered Nurses; Brickmasons and Blockmasons; Taxi Drivers; Geological Technicians; Geothermal Technicians. Examples of jobs where Achievement Orientation was rated 25 by experts: Private Household Cooks; Janitors; Paramedics; Prepress Technicians and Workers; Teaching Assistants.

Input Field	Description	O*NET Database Version	Example
			Examples of jobs where Achievement Orientation was rated 22-23 by experts: Maids and Housekeeping Cleaners; Nannies; Ophthalmic Medical Technicians; Fishing and Hunting Workers; Laundry and Dry-Cleaning Workers.

Note: For the ExampleJobTitles field, we only included reported titles with the column “Show in My Next Move” marked as “Y” in the Sample of Reported Titles table from O*NET 29.3. **For the JobTasks field, we included job tasks that were identified as core tasks. We also included emerging tasks categorized as “New” and that had a domain source of “AI/SME.” For occupations without task importance ratings, we included all job tasks for those occupations.

Table B.2. CustomFewShot Field Text for Each Work Style

Work Style	CustomFewShot
Achievement Orientation	When making your ratings, consider the following examples: Examples of jobs where Achievement Orientation was rated 35 by experts: Climate Change Policy Analysts; Door-to-Door Salespeople; Fundraising Managers. Examples of jobs where Achievement Orientation was rated 32-33 by experts: Biochemists and Biophysicists; Kindergarten through Secondary Education Administrators; Postsecondary Chemistry Teachers; Financial Risk Specialists; Penetration Testers. Examples of jobs where Achievement Orientation was rated 30 by experts: Computer Systems Analysts; Cost Estimators; Geodetic Surveyors; Soil and Plant Scientists; Water Resource Specialists. Examples of jobs where Achievement Orientation was rated 27-28 by experts: Registered Nurses; Brickmasons and Blockmasons; Taxi Drivers; Geological Technicians; Geothermal Technicians. Examples of jobs where Achievement Orientation was rated 25 by experts: Private Household Cooks; Janitors; Paramedics; Prepress Technicians and Workers; Teaching Assistants. Examples of jobs where Achievement Orientation was rated 22-23 by experts: Maids and Housekeeping Cleaners; Nannies; Ophthalmic Medical Technicians; Fishing and Hunting Workers; Laundry and Dry-Cleaning Workers.
Attention to Detail	When making your ratings, consider the following examples: Examples of jobs where Attention to Detail was rated 35 by experts: Financial Risk Specialists; Prepress Technicians and Workers; Proofreaders and Copy Markers; Registered Nurses; Statisticians. Examples of jobs where Attention to Detail was rated 32-33 by experts: Tank Car, Truck, and Ship Loaders; Craft Artists; Librarians and Media Collections Specialists; Paramedics; Purchasing Agents. Examples of jobs where Attention to Detail was rated 30 by experts: Administrative Services Managers; Animal Caretakers; First-Line Supervisors of Farming, Fishing, and Forestry Workers; Instructional Coordinators; Switchboard Operators. Examples of jobs where Attention to Detail was rated 27-28 by experts: Teaching Assistants; Textile Cutting Machine Setters, Operators, and Tenders; Wellhead Pumpers; Dishwashers; Park Naturalists. Examples of jobs where Attention to Detail was rated 25 by experts: Fishing and Hunting Workers; Poets, Lyricists, and Creative Writers; Recycling and Reclamation Workers.

Work Style	CustomFewShot
Cautiousness	When making your ratings, consider the following examples: Examples of jobs where Cautiousness was rated 35 by experts: Commercial Pilots; Petroleum Pump System Operators, Refinery Operators, and Gaugers; Security Managers. Examples of jobs where Cautiousness was rated 32-33 by experts: Registered Nurses; Roof Bolters in Mining; Textile Cutting Machine Setters, Operators, and Tenders; Aerospace Engineers; Judges, Magistrate Judges, and Magistrates. Examples of jobs where Cautiousness was rated 30 by experts: Automotive Service Technicians and Mechanics; Brickmasons and Blockmasons; Computer Numerically Controlled Tool Operators; Occupational Therapy Aides; Paramedics. Examples of jobs where Cautiousness was rated 27-28 by experts: Geological Technicians; Occupational Therapists; Recreation Workers; First-Line Supervisors of Farming, Fishing, and Forestry Workers; Geodetic Surveyors. Examples of jobs where Cautiousness was rated 25 by experts: Brokerage Clerks; Private Household Cooks; Embalmers; Farm Labor Contractors; Prepress Technicians and Workers. Examples of jobs where Cautiousness was rated 22-23 by experts: Librarians and Media Collections Specialists; Maids and Housekeeping Cleaners; Switchboard Operators; Fundraising Managers; Park Naturalists. Examples of jobs where Cautiousness was rated 20 by experts: Coaches and Scouts. Examples of jobs where Cautiousness was rated 17-18 by experts: Actors; Door-to-Door Salespeople.
Dependability	When making your ratings, consider the following examples: Examples of jobs where Dependability was rated 35 by experts: Commercial Pilots; Meeting, Convention, and Event Planners; Nannies; Proofreaders and Copy Markers; Security Managers. Examples of jobs where Dependability was rated 32-33 by experts: Animal Caretakers; Urologists; Water Resource Specialists; Farm Labor Contractors; Neuropsychologists. Examples of jobs where Dependability was rated 30 by experts: Appraisers of Personal and Business Property; Door-to-Door Salespeople; Fishing and Hunting Workers; Librarians and Media Collections Specialists; Soil and Plant Scientists. Examples of jobs where Dependability was rated 27-28 by experts: Actors. Examples of jobs where Dependability was rated 25 by experts: Craft Artists.

Work Style	CustomFewShot
Self-Confidence	When making your ratings, consider the following examples: Examples of jobs where Self-Confidence was rated 32-33 by experts: Judges, Magistrate Judges, and Magistrates; Purchasing Agents; Real Estate Sales Agents. Examples of jobs where Self-Confidence was rated 30 by experts: Financial Risk Specialists; Fundraising Managers; Mental Health and Substance Abuse Social Workers; Neuropsychologists; Urologists. Examples of jobs where Self-Confidence was rated 27-28 by experts: Financial Quantitative Analysts; Recreation Workers; Statisticians; Postsecondary Chemistry Teachers; Physician Assistants. Examples of jobs where Self-Confidence was rated 25 by experts: Automotive Service Technicians and Mechanics; Brokerage Clerks; Craft Artists; Park Naturalists; Penetration Testers. Examples of jobs where Self-Confidence was rated 22-23 by experts: Fishing and Hunting Workers; Personal Care Aides; Pest Control Workers; Dishwashers; Medical Secretaries and Administrative Assistants. Examples of jobs where Self-Confidence was rated 20 by experts: Embalmers; Geothermal Technicians.
Leadership Orientation	When making your ratings, consider the following examples: Examples of jobs where Leadership Orientation was rated 35 by experts: First-Line Supervisors of Farming, Fishing, and Forestry Workers. Examples of jobs where Leadership Orientation was rated 32-33 by experts: Judges, Magistrate Judges, and Magistrates; Security Managers; Emergency Medicine Physicians; Farm Labor Contractors. Examples of jobs where Leadership Orientation was rated 30 by experts: Administrative Services Managers; Construction and Building Inspectors; Meeting, Convention, and Event Planners; Recreation Workers. Examples of jobs where Leadership Orientation was rated 27-28 by experts: Petroleum Engineers; Sales Representatives; Set and Exhibit Designers. Examples of jobs where Leadership Orientation was rated 25 by experts: Nannies; Park Naturalists; Security Guards; Urologists. Examples of jobs where Leadership Orientation was rated 22-23 by experts: Animal Caretakers; Pest Control Workers; Statisticians; Teaching Assistants; Water Resource Specialists. Examples of jobs where Leadership Orientation was rated 20 by experts: Brokerage Clerks; Restaurant Cooks; Geothermal Technicians; Physician Assistants; Taxi Drivers. Examples of jobs where Leadership Orientation was rated 17-18 by experts: Computer Programmers; Prepress Technicians and Workers; Proofreaders and Copy Markers; Tapers; Maids and Housekeeping Cleaners.

Work Style	CustomFewShot
Social Orientation	When making your ratings, consider the following examples: Examples of jobs where Social Orientation was rated 35 by experts: Fundraising Managers; Sales Representatives. Examples of jobs where Social Orientation was rated 32-33 by experts: Actors; Kindergarten through Secondary Education Administrators; Meeting, Convention, and Event Planners; Rehabilitation Counselors; Teaching Assistants. Examples of jobs where Social Orientation was rated 30 by experts: Instructional Coordinators; Occupational Therapy Aides; Switchboard Operators. Examples of jobs where Social Orientation was rated 27-28 by experts: Postmasters and Mail Superintendents; Public Safety Telecommunicators; Purchasing Agents; Pediatric Surgeons; Taxi Drivers. Examples of jobs where Social Orientation was rated 25 by experts: Amusement and Recreation Attendants; Construction and Building Inspectors; Paralegals and Legal Assistants; Security Guards; Security Managers. Examples of jobs where Social Orientation was rated 22-23 by experts: Financial Quantitative Analysts; Water Resource Specialists; Aerospace Engineers; Embalmers; Maids and Housekeeping Cleaners. Examples of jobs where Social Orientation was rated 20 by experts: Dishwashers; Geological Technicians; Statisticians; Textile Cutting Machine Setters, Operators, and Tenders; Title Examiners, Abstractors, and Searchers. Examples of jobs where Social Orientation was rated 17-18 by experts: Fishing and Hunting Workers; Janitors; Penetration Testers; Refractory Materials Repairers; Wellhead Pumpers.
Cooperation	When making your ratings, consider the following examples: Examples of jobs where Cooperation was rated 35 by experts: Radiation Therapists. Examples of jobs where Cooperation was rated 32-33 by experts: Meeting, Convention, and Event Planners; Middle School Teachers; Kindergarten through Secondary Education Administrators; Paramedics; Public Safety Telecommunicators. Examples of jobs where Cooperation was rated 30 by experts: Ophthalmic Medical Technicians; Park Naturalists; Set and Exhibit Designers. Examples of jobs where Cooperation was rated 27-28 by experts: Door-to-Door Salespeople; Financial Quantitative Analysts; Postsecondary Chemistry Teachers; Pest Control Workers; Petroleum Engineers. Examples of jobs where Cooperation was rated 25 by experts: Brokerage Clerks; Conveyor Operators and Tenders; Farmers and Ranchers; Geothermal Technicians; Sales Representatives. Examples of jobs where Cooperation was rated 22-23 by experts: Business Intelligence Analysts; Geodetic Surveyors; Soil and Plant Scientists; Floor Sanders and Finishers; Craft Artists.

Work Style	CustomFewShot
	Examples of jobs where Cooperation was rated 20 by experts: Appraisers of Personal and Business Property; Cleaning, Washing, and Metal Pickling Equipment Operators and Tenders; Roof Bolters in Mining.
Empathy	When making your ratings, consider the following examples: Examples of jobs where Empathy was rated 35 by experts: Personal Care Aides; Radiation Therapists; Rehabilitation Counselors. Examples of jobs where Empathy was rated 32-33 by experts: Nannies; Registered Nurses; Kindergarten through Secondary Education Administrators; Teaching Assistants. Examples of jobs where Empathy was rated 30 by experts: Coaches and Scouts; Ophthalmic Medical Technicians; Police and Sheriff's Patrol Officers; Public Safety Telecommunicators. Examples of jobs where Empathy was rated 27-28 by experts: Administrative Services Managers; Meeting, Convention, and Event Planners; Pediatric Surgeons; First-Line Supervisors of Non-Retail Sales Workers; Recreation Workers. Examples of jobs where Empathy was rated 25 by experts: Animal Caretakers; Door-to-Door Salespeople; Judges, Magistrate Judges, and Magistrates; Park Naturalists; Taxi Drivers. Examples of jobs where Empathy was rated 22-23 by experts: Aerospace Engineers; Computer Programmers; Restaurant Cooks; Financial Quantitative Analysts; Water Resource Specialists. Examples of jobs where Empathy was rated 20 by experts: Cost Estimators; Proofreaders and Copy Markers; Soil and Plant Scientists; Title Examiners, Abstractors, and Searchers; Wellhead Pumpers. Examples of jobs where Empathy was rated 17-18 by experts: Non-Destructive Testing Specialists; Penetration Testers; Appraisers of Personal and Business Property; Fishing and Hunting Workers.

Work Style	CustomFewShot
Humility	When making your ratings, consider the following examples: Examples of jobs where Humility was rated 27-28 by experts: Nannies; Farm Labor Contractors; Neuropsychologists; Personal Care Aides; Rehabilitation Counselors. Examples of jobs where Humility was rated 25 by experts: Brokerage Clerks; Kindergarten through Secondary Education Administrators; Intelligence Analysts; Meeting, Convention, and Event Planners; Physician Assistants. Examples of jobs where Humility was rated 22-23 by experts: Electrical and Electronics Repairers of Commercial and Industrial Equipment; Geodetic Surveyors; Metal and Plastic Model Makers; Pest Control Workers; Taxi Drivers. Examples of jobs where Humility was rated 20 by experts: Postsecondary Chemistry Teachers; Forest and Conservation Workers; Set and Exhibit Designers; Tapers; Water Resource Specialists. Examples of jobs where Humility was rated 17-18 by experts: Financial Risk Specialists; Fundraising Managers; Security Guards; Fishing and Hunting Workers; Penetration Testers. Examples of jobs where Humility was rated 15 by experts: Emergency Medicine Physicians. Examples of jobs where Humility was rated 12-13 by experts: Sales Representatives. Examples of jobs where Humility was rated 10 by experts: Actors.
Integrity	When making your ratings, consider the following examples: Examples of jobs where Integrity was rated 35 by experts: Intelligence Analysts; Security Managers. Examples of jobs where Integrity was rated 32-33 by experts: Brokerage Clerks; Aerospace Engineers; Airline Pilots, Copilots, and Flight Engineers; Mental Health and Substance Abuse Social Workers; Pediatric Surgeons. Examples of jobs where Integrity was rated 30 by experts: Coaches and Scouts; Farm Labor Contractors; First-Line Supervisors of Non-Retail Sales Workers; Instructional Coordinators; Roof Bolters in Mining. Examples of jobs where Integrity was rated 27-28 by experts: Administrative Services Managers; Climate Change Policy Analysts; Computer Programmers; Door-to-Door Salespeople; Ophthalmic Medical Technicians. Examples of jobs where Integrity was rated 25 by experts: Geothermal Technicians; Librarians and Media Collections Specialists; Proofreaders and Copy Markers; Sailors and Marine Oilers; Sales Representatives. Examples of jobs where Integrity was rated 22-23 by experts: Dishwashers; Prepress Technicians and Workers; Actors.

Work Style	CustomFewShot
Sincerity	When making your ratings, consider the following examples: Examples of jobs where Sincerity was rated 35 by experts: Mental Health and Substance Abuse Social Workers. Examples of jobs where Sincerity was rated 32-33 by experts: Radiation Therapists; Personal Care Aides. Examples of jobs where Sincerity was rated 30 by experts: Judges, Magistrate Judges, and Magistrates; Nannies; Occupational Therapy Aides; Physician Assistants; Teaching Assistants. Examples of jobs where Sincerity was rated 27-28 by experts: Meeting, Convention, and Event Planners; Urologists; Door-to-Door Salespeople; Administrative Services Managers; Instructional Coordinators. Examples of jobs where Sincerity was rated 25 by experts: Automotive Service Technicians and Mechanics; Postsecondary Chemistry Teachers; First-Line Supervisors of Farming, Fishing, and Forestry Workers; Paramedics; Park Naturalists. Examples of jobs where Sincerity was rated 22-23 by experts: Brickmasons and Blockmasons; Restaurant Cooks; Maids and Housekeeping Cleaners; Sailors and Marine Oilers; Statisticians. Examples of jobs where Sincerity was rated 20 by experts: Craft Artists; Government Property Inspectors and Investigators; Laundry and Dry-Cleaning Workers; Roof Bolters in Mining; Wellhead Pumpers.
Self-Control	When making your ratings, consider the following examples: Examples of jobs where Self-Control was rated 35 by experts: Commercial Pilots; Paramedics; Public Safety Telecommunicators. Examples of jobs where Self--Control was rated 32-33 by experts: Kindergarten through Secondary Education Administrators; Mental Health and Substance Abuse Social Workers; Meeting, Convention, and Event Planners; Nannies; Registered Nurses. Examples of jobs where Self-Control was rated 30 by experts: Restaurant Cooks; First-Line Supervisors of Non-Retail Sales Workers; Intelligence Analysts; Rehabilitation Counselors; Sailors and Marine Oilers. Examples of jobs where Self-Control was rated 27-28 by experts: Postsecondary Chemistry Teachers; Door-to-Door Salespeople; Neuropsychologists; Computer, Automated Teller, and Office Machine Repairers; Production, Planning, and Expediting Clerks. Examples of jobs where Self--Control was rated 25 by experts: Biochemists and Biophysicists; Librarians and Media Collections Specialists; Maids and Housekeeping Cleaners; Occupational Therapy Aides; Statisticians. Examples of jobs where Self-Control was rated 22-23 by experts: Computer Programmers; Conveyor Operators and Tenders; Janitors; Laundry and Dry-Cleaning Workers; Tapers.

Work Style	CustomFewShot
	Examples of jobs where Self-Control was rated 20 by experts: Poets, Lyricists, and Creative Writers.
Stress Tolerance	When making your ratings, consider the following examples: Examples of jobs where Stress Tolerance was rated 35 by experts: Paramedics; Police and Sheriff's Patrol Officers; Public Safety Telecommunicators. Examples of jobs where Stress Tolerance was rated 32-33 by experts: Meeting, Convention, and Event Planners; Registered Nurses; Middle School Teachers; Security Managers. Examples of jobs where Stress Tolerance was rated 30 by experts: Restaurant Cooks; First-Line Supervisors of Farming, Fishing, and Forestry Workers; Intelligence Analysts; Rehabilitation Counselors; Urologists. Examples of jobs where Stress Tolerance was rated 27-28 by experts: Penetration Testers; Production, Planning, and Expediting Clerks; Recreation Workers; Aerospace Engineers; Postmasters and Mail Superintendents. Examples of jobs where Stress Tolerance was rated 25 by experts: Brokerage Clerks; Fundraising Managers; Geothermal Technicians; Maids and Housekeeping Cleaners; Wind Energy Engineers. Examples of jobs where Stress Tolerance was rated 22-23 by experts: Craft Artists; Fishing and Hunting Workers; Geological Technicians; Appraisers of Personal and Business Property; Prepress Technicians and Workers.
Adaptability	When making your ratings, consider the following examples: Examples of jobs where Adaptability was rated 32-33 by experts: Instructional Coordinators; Police and Sheriff's Patrol Officers; Wind Energy Engineers; Emergency Medicine Physicians. Examples of jobs where Adaptability was rated 30 by experts: Airline Pilots, Copilots, and Flight Engineers; Postsecondary Chemistry Teachers; Coaches and Scouts; Kindergarten through Secondary Education Administrators; Production, Planning, and Expediting Clerks. Examples of jobs where Adaptability was rated 27-28 by experts: First-Line Supervisors of Farming, Fishing, and Forestry Workers; Recreation Workers; Rehabilitation Counselors; Statisticians; Urologists. Examples of jobs where Adaptability was rated 25 by experts: Computer, Automated Teller, and Office Machine Repairers; Personal Care Aides; Pest Control Workers; Poets, Lyricists, and Creative Writers; Security Guards. Examples of jobs where Adaptability was rated 22-23 by experts: Animal Caretakers; Proofreaders and Copy Markers; Geodetic Surveyors; Laundry and Dry-Cleaning Workers; Tapers. Examples of jobs where Adaptability was rated 20 by experts: Amusement and Recreation Attendants; Refractory Materials Repairers; Roof Bolters in Mining; Tank Car, Truck, and Ship Loaders; Title Examiners, Abstractors, and Searchers.

Work Style	CustomFewShot
Innovation	When making your ratings, consider the following examples: Examples of jobs where Innovation was rated 32-33 by experts: Craft Artists; Soil and Plant Scientists; Instructional Coordinators; Wind Energy Engineers; Penetration Testers. Examples of jobs where Innovation was rated 30 by experts: Financial Quantitative Analysts; Urologists; Water Resource Specialists. Examples of jobs where Innovation was rated 27-28 by experts: Fundraising Managers; Neuropsychologists; Business Intelligence Analysts; Middle School Teachers; Occupational Therapists. Examples of jobs where Innovation was rated 25 by experts: Climate Change Policy Analysts; Electrical and Electronics Drafters; Librarians and Media Collections Specialists; Metal and Plastic Model Makers; Purchasing Agents. Examples of jobs where Innovation was rated 22-23 by experts: Glaziers; Nannies; Fishing and Hunting Workers; Non-Destructive Testing Specialists; Farm Labor Contractors. Examples of jobs where Innovation was rated 20 by experts: Brokerage Clerks; First-Line Supervisors of Farming, Fishing, and Forestry Workers; Government Property Inspectors and Investigators; Maids and Housekeeping Cleaners; Personal Care Aides. Examples of jobs where Innovation was rated 17-18 by experts: Petroleum Pump System Operators, Refinery Operators, and Gaugers; Public Safety Telecommunicators; Security Guards; Janitors; Roof Bolters in Mining. Examples of jobs where Innovation was rated 15 by experts: Amusement and Recreation Attendants; Semiconductor Processing Technicians.

Work Style	CustomFewShot
Intellectual Curiosity	When making your ratings, consider the following examples: Examples of jobs where Intellectual Curiosity was rated 35 by experts: Biochemists and Biophysicists; Financial Risk Specialists. Examples of jobs where Intellectual Curiosity was rated 32-33 by experts: Financial Quantitative Analysts; Penetration Testers; Aerospace Engineers; Intelligence Analysts; Librarians and Media Collections Specialists. Examples of jobs where Intellectual Curiosity was rated 30 by experts: Business Intelligence Analysts; Commercial Pilots; Non-Destructive Testing Specialists; Paramedics; Set and Exhibit Designers. Examples of jobs where Intellectual Curiosity was rated 27-28 by experts: Automotive Service Technicians and Mechanics; Craft Artists; Appraisers of Personal and Business Property; Private Household Cooks; Farmers and Ranchers. Examples of jobs where Intellectual Curiosity was rated 25 by experts: Restaurant Cooks; First-Line Supervisors of Farming, Fishing, and Forestry Workers; Police and Sheriff's Patrol Officers; Proofreaders and Copy Markers; Rehabilitation Counselors. Examples of jobs where Intellectual Curiosity was rated 22-23 by experts: Brokerage Clerks; Fishing and Hunting Workers; Glaziers; Ophthalmic Medical Technicians; Petroleum Pump System Operators, Refinery Operators, and Gaugers. Examples of jobs where Intellectual Curiosity was rated 20 by experts: Adhesive Bonding Machine Operators and Tenders; Brickmasons and Blockmasons; Personal Care Aides; Switchboard Operators; Taxi Drivers. Examples of jobs where Intellectual Curiosity was rated 17-18 by experts: Maids and Housekeeping Cleaners; Amusement and Recreation Attendants; Dishwashers; Tapers.

Work Style	CustomFewShot
Tolerance for Ambiguity	When making your ratings, consider the following examples: Examples of jobs where Tolerance for Ambiguity was rated 32-33 by experts: Intelligence Analysts. Examples of jobs where Tolerance for Ambiguity was rated 30 by experts: Computer Programmers; Door-to-Door Salespeople; Emergency Medicine Physicians; Fundraising Managers; Middle School Teachers. Examples of jobs where Tolerance for Ambiguity was rated 27-28 by experts: Public Safety Telecommunicators; Rehabilitation Counselors; Set and Exhibit Designers; Biochemists and Biophysicists; Judges, Magistrate Judges, and Magistrates. Examples of jobs where Tolerance for Ambiguity was rated 25 by experts: Administrative Services Managers; Production, Planning, and Expediting Clerks; Sailors and Marine Oilers; Urologists; Paralegals and Legal Assistants. Examples of jobs where Tolerance for Ambiguity was rated 22-23 by experts: Biofuels Production Managers; Cost Estimators; Park Naturalists; Instructional Coordinators; Forest and Conservation Workers. Examples of jobs where Tolerance for Ambiguity was rated 20 by experts: Brokerage Clerks; Computer, Automated Teller, and Office Machine Repairers; Conveyor Operators and Tenders; Government Property Inspectors and Investigators; Prepress Technicians and Workers. Examples of jobs where Tolerance for Ambiguity was rated 17-18 by experts: Geothermal Technicians; Radiation Therapists; Brickmasons and Blockmasons; Textile Cutting Machine Setters, Operators, and Tenders; Wellhead Pumpers. Examples of jobs where Tolerance for Ambiguity was rated 15 by experts: Cleaning, Washing, and Metal Pickling Equipment Operators and Tenders; Tapers.
Initiative	When making your ratings, consider the following examples: Examples of jobs where Initiative was rated 32-33 by experts: Penetration Testers; Real Estate Sales Agents. Examples of jobs where Initiative was rated 30 by experts: Private Household Cooks; Farm Labor Contractors; Instructional Coordinators; Nannies; Registered Nurses. Examples of jobs where Initiative was rated 27-28 by experts: Financial Quantitative Analysts; Occupational Therapists; First-Line Supervisors of Farming, Fishing, and Forestry Workers; Soil and Plant Scientists; Statisticians. Examples of jobs where Initiative was rated 25 by experts: Automotive Service Technicians and Mechanics; Forest and Conservation Workers; Geodetic Surveyors; Maids and Housekeeping Cleaners; Paramedics. Examples of jobs where Initiative was rated 22-23 by experts: Dishwashers; Electrical and Electronics Repairers of Commercial and Industrial Equipment; Glaziers; Prepress Technicians and Workers; Electrical and Electronics Drafters.

Work Style	CustomFewShot
	Examples of jobs where Initiative was rated 20 by experts: Appraisers of Personal and Business Property; Tapers; Wellhead Pumpers. Examples of jobs where Initiative was rated 17-18 by experts: Brickmasons and Blockmasons.
Optimism	When making your ratings, consider the following examples: Examples of jobs where Optimism was rated 32-33 by experts: Personal Care Aides; Mental Health and Substance Abuse Social Workers; Nannies; Teaching Assistants. Examples of jobs where Optimism was rated 30 by experts: Coaches and Scouts; Meeting, Convention, and Event Planners; Occupational Therapy Aides; Recreation Workers; Registered Nurses. Examples of jobs where Optimism was rated 27-28 by experts: Commercial Pilots; First-Line Supervisors of Farming, Fishing, and Forestry Workers; First-Line Supervisors of Non-Retail Sales Workers; Physician Assistants; Police and Sheriff's Patrol Officers. Examples of jobs where Optimism was rated 25 by experts: Administrative Services Managers; Emergency Medicine Physicians; Instructional Coordinators; Judges, Magistrate Judges, and Magistrates; Neuropsychologists. Examples of jobs where Optimism was rated 22-23 by experts: Computer Systems Analysts; Computer, Automated Teller, and Office Machine Repairers; Aerospace Engineers; Financial Quantitative Analysts; Set and Exhibit Designers. Examples of jobs where Optimism was rated 20 by experts: Business Intelligence Analysts; Forest and Conservation Workers; Geodetic Surveyors; Glaziers; Soil and Plant Scientists.

Work Style	CustomFewShot
Perseverance	When making your ratings, consider the following examples: Examples of jobs where Perseverance was rated 35 by experts: Sales Representatives. Examples of jobs where Perseverance was rated 32-33 by experts: Airline Pilots, Copilots, and Flight Engineers; Real Estate Sales Agents; Coaches and Scouts; Fundraising Managers; Police and Sheriff's Patrol Officers. Examples of jobs where Perseverance was rated 30 by experts: Biochemists and Biophysicists; Postsecondary Chemistry Teachers; Commercial Pilots; First-Line Supervisors of Non-Retail Sales Workers; Mental Health and Substance Abuse Social Workers. Examples of jobs where Perseverance was rated 27-28 by experts: Instructional Coordinators; Poets, Lyricists and Creative Writers; Public Safety Telecommunicators; Glaziers; Water Resource Specialists. Examples of jobs where Perseverance was rated 25 by experts: Computer Numerically Controlled Tool Operators; Private Household Cooks; Radiation Therapists; Refractory Materials Repairers; Tapers. Examples of jobs where Perseverance was rated 22-23 by experts: Adhesive Bonding Machine Operators and Tenders; Conveyor Operators and Tenders; Janitors; Park Naturalists; Switchboard Operators.

Appendix C: Summary of Process for Generating Final Hybrid Generative AI-Expert Ratings

This appendix outlines the 6-step process used to generate a final hybrid Generative AI (genAI) expert rating for a Work Style in a given occupation.

Step 1: The prompt template described in Appendix B is populated with information for the given Work Style-occupation combination to be rated and subsequently run through three LLMs (Claude 3.5 Sonnet v1, Claude 3.5 Sonnet v2, and Llama 3.3 70B Instruct) on three separate occasions, yielding a total of nine impact ratings for the given Work Style-occupation combination. For all runs, hyperparameter settings are. 0.2 for temperature, and 0.2 for top p.

Step 2: The nine ratings obtained in Step 1 are rescaled on a 5-35 point scale (as specified in the prompt in Appendix B). These ratings are converted to the -3 to +3 impact rating scale metric by dividing by 5 and then adding 4. Next, the three rescaled ratings for each model are averaged together to form a raw composite rating based on each LLM. These composites are subsequently averaged together to form a raw Work Style rating for the occupation being rated.

Step 3: The following transformation is run on the raw Work Style rating:

Transformed WS Rating = $(\text{Raw WS Rating} + 4)^{x}$

The value of the exponent (x) used for this transformation varies by Work Style and was determined by identifying the value that minimized the difference in skew between the transformed Work Style rating distribution and expert rating distribution across the sample of 125 O*NET 29.3 occupations used for evaluation. Table C.1 lists the transformation exponent used for each Work Style to produce a transformed Work Style rating.

Table C.1. Exponents Used to Calculate Transformed Work Style Ratings

Work Style	Transformation Exponent
Achievement Orientation	4
Adaptability	4
Attention To Detail	4
Cautiousness	3
Cooperation	4
Dependability	3
Empathy	3
Humility	3
Initiative	4
Innovation	2
Integrity	4
Intellectual Curiosity	4
Leadership Orientation	2
Optimism	4
Perseverance	4
Self Confidence	4

Work Style	Transformation Exponent
Self Control	4
Sincerity	4
Social Orientation	n/a*
Stress Tolerance	4
Tolerance For Ambiguity	n/a*

Note. *For Social Orientation and Tolerance for Ambiguity, no transformation was done as the skew of the raw distribution was similar to the skew of the experts. Thus, for Social Orientation and Tolerance for Ambiguity, the raw Work Style rating is effectively used as the transformed Work Style rating in subsequent steps.

Step 4: The transformed Work Style ratings resulting from Step 3 are converted to z-scores using means and Standard Deviations (*SD*s) of the transformed Work Style ratings in the evaluation sample of 125 occupations using the following formula

$$\textbf{Transformed WS Rating Z-Score} = \frac{(Transformed\ WS\ Rating - Mean\ of\ Transformed\ WS\ Rating)}{SD\ of\ Transformed\ WS\ Rating}$$

Table C.2. lists the means and *SD*s used for each Work Style to calculate the Z-scores.

Table C.2. Means and Standard Deviations Used to Calculate Transformed Work Style Rating Z-scores

Work Style	Mean of Transformed WS Rating	*SD* of Transformed WS Rating
Achievement Orientation	129.37223	495.32007
Adaptability	116.94776	427.99042
Attention To Detail	1739.09124	459.07219
Cautiousness	224.05968	62.53354
Cooperation	1221.80278	48.27074
Dependability	289.85596	35.35235
Empathy	145.13779	8.14386
Humility	117.60407	39.41349
Initiative	1138.84472	445.87416
Innovation	25.76203	9.45011
Integrity	1532.93559	492.22890
Intellectual Curiosity	1141.90685	613.27184
Leadership Orientation	25.38898	9.66872
Optimism	831.32546	432.63405
Perseverance	134.93132	37.55814
Self Confidence	1159.69338	41.17596
Self Control	1314.45493	459.80449
Sincerity	881.61294	492.51250
Social Orientation	1.11804	.95014
Stress Tolerance	1281.12159	435.08325
Tolerance For Ambiguity	1.06916	.92663

Note. n = 125 (occupations in evaluation sample). Data used as inputs to the prompt template underlying these results were drawn from O*NET 29.3.

Step 5: The transformed Work Style rating z-scores from Step 4 are re-scaled using the following formula, which effectively treats the expert ratings from the evaluation sample of 125 occupations as a norming (calibration) sample.

Final Unrounded WS Rating = Mean of Expert WS Rating + (Transformed WS Rating *Z*-Score x *SD* of Expert WS Rating)

Table C.3 lists the means and *SD*s of expert ratings used for each Work Style to calculate final unrounded Work Style ratings referenced above.

Table C.3. Means and Standard Deviations of Expert Work Style Ratings Used to Calculate Final Unrounded Work Style Ratings

Work Style	Mean of Expert WS Rating	*SD* of Expert WS Rating
Achievement Orientation	1.43800	.65594
Adaptability	1.18520	.72016
Attention To Detail	2.27040	.57614
Cautiousness	1.66920	.82386
Cooperation	1.31640	.78890
Dependability	2.40880	.34624
Empathy	.71800	.92781
Humility	.30520	.62025
Initiative	1.03200	.65314
Innovation	.70720	1.00911
Integrity	1.68200	.70846
Intellectual Curiosity	1.09720	.99702
Leadership Orientation	.65560	1.02476
Optimism	.70160	.80043
Perseverance	1.46520	.55791
Self Confidence	1.16920	.63860
Self Control	1.34040	.72177
Sincerity	.82080	.67074
Social Orientation	.88840	.96937
Stress Tolerance	1.41040	.64092
Tolerance For Ambiguity	.64280	.86739

Note. n = 125 (occupations in evaluation sample).

Step 6: To arrive at the final Work Style rating for a given Work Style-occupation combination, final unrounded ratings resulting from Step 5 that are less than -3.0 are rounded up to -3.0, and ratings greater than +3.0 are rounded down to +3.

Appendix D: Work Style Intercorrelations based on LLM Ratings and Analyst Ratings for the Evaluation Sample of Occupations (*n* = 125)

Table D.1. Work Style Intercorrelations based on LLM Ratings

	Work Style	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Achievement Orientation	1.00																				
2	Adaptability	.65	1.00																			
3	Attention To Detail	.31	.16	1.00																		
4	Cautiousness	.01	-.06	.62	1.00																	
5	Cooperation	.24	.57	-.05	-.14	1.00																
6	Dependability	.16	.30	.43	.50	.43	1.00															
7	Empathy	.25	.48	-.07	-.23	.87	.29	1.00														
8	Humility	-.07	.15	.12	.11	.58	.39	.56	1.00													
9	Initiative	.74	.78	.04	-.19	.53	.19	.51	.15	1.00												
10	Innovation	.71	.66	.15	-.18	.22	-.12	.24	.04	.71	1.00											
11	Integrity	.49	.46	.41	.39	.45	.63	.45	.34	.50	.24	1.00										
12	Intellectual Curiosity	.77	.62	.49	.18	.24	.18	.23	.13	.63	.80	.50	1.00									
13	Leadership Orientation	.66	.62	.06	-.02	.50	.43	.44	.27	.72	.40	.55	.46	1.00								
14	Optimism	.18	.43	-.39	-.43	.76	.13	.82	.40	.50	.19	.25	.02	.41	1.00							
15	Perseverance	.69	.65	.19	.07	.36	.35	.36	-.03	.66	.47	.56	.52	.61	.34	1.00						
16	Self Confidence	.76	.70	.16	-.01	.47	.30	.43	.07	.68	.50	.51	.55	.77	.40	.71	1.00					
17	Self Control	.33	.55	.15	.19	.69	.66	.63	.34	.47	.06	.68	.23	.62	.52	.56	.60	1.00				
18	Sincerity	.31	.47	.03	-.09	.82	.40	.88	.60	.53	.21	.58	.29	.54	.72	.42	.48	.68	1.00			
19	Social Orientation	.43	.60	-.15	-.36	.81	.24	.84	.41	.66	.36	.42	.28	.61	.85	.45	.59	.59	.78	1.00		
20	Stress Tolerance	.33	.54	.16	.24	.62	.62	.58	.26	.44	.08	.59	.27	.54	.43	.57	.61	.86	.59	.49	1.00	
21	Tolerance For Ambiguity	.74	.78	.12	-.11	.42	.10	.42	.11	.81	.77	.52	.73	.58	.38	.71	.68	.40	.46	.53	.42	1.00

Note. *N* = 125 (unit of analysis is occupation). Data used as inputs to the prompt template underlying these results were drawn from O*NET 29.3.

Table D.2. Work Style Intercorrelations based on Analyst Ratings

Work Style		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Achievement Orientation	1.00																				
2	Adaptability	.66	1.00																			
3	Attention To Detail	.29	.18	1.00																		
4	Cautiousness	-.13	.02	.55	1.00																	
5	Cooperation	.29	.43	-.15	-.18	1.00																
6	Dependability	.10	.35	.31	.50	.49	1.00															
7	Empathy	.25	.36	.00	.01	.79	.54	1.00														
8	Humility	.17	.37	-.04	.02	.77	.48	.87	1.00													
9	Initiative	.74	.57	.09	-.17	.45	.18	.38	.32	1.00												
10	Innovation	.65	.66	.08	-.22	.13	-.15	.07	.13	.53	1.00											
11	Integrity	.43	.47	.38	.20	.52	.63	.55	.47	.38	.14	1.00										
12	Intellectual Curiosity	.75	.61	.45	.06	.20	.19	.21	.22	.64	.67	.45	1.00									
13	Leadership Orientation	.62	.50	.08	.02	.54	.41	.45	.32	.54	.26	.50	.39	1.00								
14	Optimism	.43	.45	-.20	-.23	.80	.37	.81	.75	.48	.21	.47	.24	.53	1.00							
15	Perseverance	.43	.53	.19	.21	.44	.49	.42	.38	.48	.30	.37	.40	.49	.45	1.00						
16	Self Confidence	.75	.71	.34	.05	.49	.44	.50	.39	.56	.45	.66	.68	.60	.58	.54	1.00					
17	Self Control	.26	.44	.02	.08	.71	.63	.74	.65	.29	.00	.65	.20	.55	.73	.53	.60	1.00				
18	Sincerity	.36	.40	.07	-.03	.80	.53	.87	.79	.47	.14	.69	.35	.50	.79	.38	.62	.76	1.00			
19	Social Orientation	.37	.36	-.32	-.27	.83	.26	.70	.69	.49	.17	.32	.22	.56	.84	.34	.45	.60	.72	1.00		
20	Stress Tolerance	.23	.48	.12	.25	.52	.69	.55	.44	.23	-.02	.62	.18	.58	.51	.60	.59	.85	.55	.36	1.00	
21	Tolerance For Ambiguity	.59	.70	.12	-.05	.37	.28	.33	.32	.46	.51	.47	.54	.42	.42	.55	.69	.42	.41	.30	.51	1.00

Note. *N* = 125 (unit of analysis is occupation).