FR-04-71

HumRRO
Human Resources Research Organization

O*NET Analyst Occupational Abilities Ratings: Analysis Cycle 2 Results

Carrie L. Noble
Suzanne Tsacoumis

Prepared for: National Center for O*NET Development
700 Wade Avenue
Raleigh, NC 27605

November 2004

Table of Contents

List of Tables

Introduction

The Occupational Information Network (O*NET) is a comprehensive system developed by the U.S. Department of Labor that provides information about nearly 1,000 occupations within the U.S. economy. The National Center for O*NET Development is in the process of collecting occupational data for over 900 occupations. The data collection effort includes job incumbent ratings on occupational tasks, skills, generalized work activities (GWA), knowledge, education and training, work styles, and work context areas. Importance and level information regarding the abilities associated with these occupations is being collected from analysts. It should be noted that there are theoretical or philosophical reasons for preferring one rater group to the other for collecting different types of data. For example, incumbents are generally more familiar with the day-to-day duties of their job, therefore they are the best source of information regarding tasks and GWAs. In contrast, it's likely that trained analysts understand the ability constructs better than incumbents and therefore should provide the ability data. Abilities are "… relatively enduring attributes of an individual's capability for performing a particular range of different tasks" (Fleishman, Costanza, & Marshall-Mies, 1999, p. 175). Abilities are sometimes referred to as traits as they tend to remain stable over long periods of time. The 52 O*NET abilities cover performance applicable to a broad range of jobs in the world's economy. These abilities are grouped into four categories: cognitive, psychomotor, physical, and sensory-perceptual constructs.

To facilitate the ability rating process, analysts are provided relevant occupational information. Trained analysts are responsible for rating the importance and level of the 52 abilities for each of the O*NET occupations. More specifically, eight trained analysts provided ratings for each occupation. For a description of the entire analyst data collection process, including the preparation and distribution of the occupational data, the steps associated with the ratings process, and the collection and management of the ability ratings, see *O*NET Analyst Occupational Abilities Ratings: Procedures* (Donsbach, Tsacoumis, Sager, & Updegraff, 2003).

To ensure a controlled data collection and management process, occupational data is being collected in groups or "analysis cycles." This report describes the results from the data collection process for the second analysis cycle of 126 occupations. Results for ratings collected in Cycle 1 are presented in Noble, Sager, Tsacoumis, Updegraff, & Donsbach (2003) and future results will be reported in separate subsequent reports. For a description of the O*NET Data Collection Publication Schedule see www.onetcenter.org. The Standard Occupational Classification Codes and Titles for the O*NET occupations included in Analysis Cycle 2 are presented in Appendix A.

Evaluation of Cycle 2 Analyst Ratings

As mentioned above, analysts provided ratings on importance and level of the 52 abilities for each of the 126 occupations in Cycle 2. The mean, standard deviation and SE_M of the importance and level ratings were computed. These results are presented in Appendix B.

Three sets of analyses were performed to evaluate the ratings that analysts provided. First, we focused on identifying the data that may be difficult to interpret based on limited agreement among raters or because there is an indication that the ability level rating is not relevant for a specific occupation. Thus, a set of recommended criteria was established which flagged: (a) an ability level rating as not relevant to an occupation because of low importance ratings, (b) an ability with too little agreement in importance ratings across raters for a particular occupation, and (c) an ability with too little agreement in level ratings across raters for a particular occupation.

The remaining three sets of analyses focused on computing measures of interrater agreement and interrater reliability. Poor agreement or reliability estimates may be an indication that there is confusion about the ability constructs, potentially due to either the nature of the definition or rater training. Specifically, the second analysis involved computing the interrater agreement among the eight raters in each rating group. Next, the interrater reliability of the raters was computed to determine the extent to which raters agreed about the order of and relative distance between constructs on a particular scale within a particular occupation. That is, this analysis provides information regarding the consistency across raters in terms of how they rate the relative importance of the 52 ability constructs to performance in a particular occupation. Finally, another interrater reliability estimate was computed to examine the consistency of ratings across occupations within constructs. In other words, this type of interrater reliability focused on the extent to which raters agree about the order of and relative distance between occupations on a particular scale for a particular construct.

Cycle 2 Recommended Data Flags

Three distinct criteria were established to flag the ability data. All three flags affect the presentation of data within the publicly available O*NET Online (online.onetcenter.org). First, the level rating of an ability was flagged as not relevant for a particular occupation if two or fewer of the eight analysts rated its importance as 2 or greater. Thus, the level rating of an ability is considered not relevant when that ability is not important for the performance of the particular occupation. In this cycle, there were 1,490 not relevant flags (see Table 1). To facilitate interpretation of these results, it should be noted that there are a total of 6,552 sets of ratings (126 occupations x 52 abilities) in the current cycle. Given this, 22.74% (1,490/6,552) of the ability ratings were flagged as not relevant. As can be noted in Table 1, these results are comparable to the findings from Cycle 1 where 23.36% of the ability level ratings were flagged as not relevant. In Cycle 2, the most common abilities identified as not relevant were Explosive Strength, Dynamic Flexibility, Night Vision, Sound Localization, and Glare Sensitivity. Once again, this is consistent with the Cycle 1 results. Given that these constructs capture fairly specific physical capabilities intuitively not required for many occupations, these results are not surprising.

Table 1. Number of Times Ability Level Flagged as Not Relevant

Element Name	Cycle 1	Cycle 2
Oral Comprehension	0	0
Written Comprehension	0	0
Oral Expression	0	0
Written Expression	0	0
Fluency of Ideas	0	2
Originality	0	7

Continued on next page

Table 1. Number of Times Ability Level Flagged as Not Relevant		
Element Name	**Cycle 1**	**Cycle 2**
Problem Sensitivity	0	0
Deductive Reasoning	0	0
Inductive Reasoning	0	0
Information Ordering	0	0
Category Flexibility	0	0
Mathematical Reasoning	0	6
Number Facility	3	5
Memorization	0	1
Speed of Closure	0	2
Flexibility of Closure	0	2
Perceptual Speed	0	1
Spatial Orientation	36	48
Visualization	0	6
Selective Attention	0	0
Time Sharing	0	0
Arm-Hand Steadiness	9	14
Manual Dexterity	9	19
Finger Dexterity	0	6
Control Precision	6	19
Multilimb Coordination	13	31
Response Orientation	30	72
Rate Control	35	88
Reaction Time	27	65
Wrist-Finger Speed	26	50
Speed of Limb Movement	28	57
Static Strength	21	38
Explosive Strength	44	104
Dynamic Strength	28	61
Trunk Strength	8	16
Stamina	21	42
Extent Flexibility	22	47
Dynamic Flexibility	52	104
Gross Body Coordination	21	46
Gross Body Equilibrium	27	67
Near Vision	0	0
Far Vision	0	4
Visual Color Discrimination	2	18
Night Vision	44	99
Peripheral Vision	44	85

Continued on next page

Table 1. Number of Times Ability Level Flagged as Not Relevant

Element Name	Cycle 1	Cycle 2
Depth Perception	11	21
Glare Sensitivity	41	93
Hearing Sensitivity	2	39
Auditory Attention	2	10
Sound Localization	44	95
Speech Recognition	0	0
Speech Clarity	0	0
Total Flags out of all possible rating	23.36% (656/2808)	22.74% (1,490/6,552)

The remaining two criteria involve the recommended suppression of identifying any ability mean or level importance rating that had a standard error of the mean (*SE_M)* greater than .51. These criteria were established to capture those ratings deemed to have insufficient agreement across raters. The value of .51 was selected because 1.0/1.96 = .51. An SE_M greater than .51 means that the upper and lower bounds of the confidence interval are more than 1 scale point away from the observed mean. The results of these two suppression criteria are presented in Table 2. As can be noted, there were only six instances (i.e., Flexibility of Closure (2), Spatial Orientation (1), Speed of Limb Movement (1), Explosive Strength (1), and Sound Localization (1)) where the mean importance rating was flagged for insufficient agreement. There were 387 insufficient agreement flags for level ratings, 13 of which were also flagged as not relevant (3% of 387). As can be noted, in Cycle 1 the mean importance and level ratings were flagged for insufficient agreement one and 157 times, respectively. For Cycle 2, the abilities that were flagged the most for the level criteria included: Wrist-Finger Speed (n=33), Speed of Closure (n=32), Flexibility of Closure (n=29), and Finger Dexterity (n=20). In many cases, the abilities with the most flags in Cycle 2 also received many flags in Cycle 1. One notable difference is Finger Dexterity which was flagged 20 times in Cycle 2 but not once in Cycle 1.

Table 2. Ability Flags Due to Large SE_M

Element Name	Frequency SE_M Importance > .51		Frequency SE_M Level > .51	
	Cycle 1	Cycle 2	Cycle 1	Cycle 2
Oral Comprehension	0	0	0	0
Written Comprehension	0	0	0	0
Oral Expression	0	0	0	0
Written Expression	0	0	0	0
Fluency of Ideas	0	0	4	11
Originality	0	0	1	3
Problem Sensitivity	0	0	0	0
Deductive Reasoning	0	0	0	0
Inductive Reasoning	0	0	0	1
Information Ordering	0	0	0	1

Continued on next page

Table 2. Ability Flags Due to Large SE_M

Element Name	Frequency SE_M Importance > .51		Frequency SE_M Level > .51	
	Cycle 1	Cycle 2	Cycle 1	Cycle 2
Category Flexibility	0	0	0	2
Mathematical Reasoning	0	0	1	7
Number Facility	0	0	1	15
Memorization	0	0	3	18
Speed of Closure	0	0	4	32
Flexibility of Closure	0	2	14	29
Perceptual Speed	0	0	12	15
Spatial Orientation	0	1	1	9
Visualization	0	0	13	19
Selective Attention	0	0	0	2
Time Sharing	0	0	0	6
Arm-Hand Steadiness	0	0	3	2
Manual Dexterity	0	0	6	8
Finger Dexterity	0	0	0	20
Control Precision	0	0	4	5
Multilimb Coordination	0	0	0	8
Response Orientation	0	0	6	8
Rate Control	0	0	3	2
Reaction Time	0	0	6	19
Wrist-Finger Speed	1	0	21	33
Speed of Limb Movement	0	1	1	4
Static Strength	0	0	4	6
Explosive Strength	0	1	3	3
Dynamic Strength	0	0	4	7
Trunk Strength	0	0	2	1
Stamina	0	0	2	3
Extent Flexibility	0	0	1	13
Dynamic Flexibility	0	0	0	3
Gross Body Coordination	0	0	0	0
Gross Body Equilibrium	0	0	4	0
Near Vision	0	0	0	0
Far Vision	0	0	16	14
Visual Color Discrimination	0	0	5	16
Night Vision	0	0	3	4
Peripheral Vision	0	0	1	2
Depth Perception	0	0	1	0
Glare Sensitivity	0	0	2	2

Continued on next page

Table 2. Ability Flags Due to Large SE_M				
	Frequency SE_M Importance > .51		Frequency SE_M Level > .51	
Element Name	Cycle 1	Cycle 2	Cycle 1	Cycle 2
Hearing Sensitivity	0	0	3	6
Auditory Attention	0	0	1	9
Sound Localization	0	1	1	9
Speech Recognition	0	0	0	8
Speech Clarity	0	0	0	2
TOTAL	0% (1/2808)	0% (6/6552)	5.59% (157/2808)	5.91% (387/6552)

Although the frequency of flagging an ability level rating was higher than the importance rating, it should be noted that the total number of level flags only reflected 5.91% of the 6,552 total ratings which is comparable to the 5.59% flagged in Cycle 1. These findings suggest a high level of agreement among the analysts. However, it may be prudent to provide additional training on the elements with the highest number of flags, particularly on Flexibility of Closure, Spatial Orientation, Speed of Limb Movement, Explosive Strength, and Sound Localization since they were flagged for both importance and level.

The detailed results of the recommended data flags and suppression criteria are depicted by the shaded cells in the results presented in Appendix B.

Cycle 2 Interrater Agreement

Interrater agreement was computed to examine the level of absolute agreement among the analysts in ratings within a construct for a particular occupation. For example, these indices identified the extent to which eight raters provided the same rating regarding the level of the ability *Written Comprehension* required to perform a particular occupation. To look at the agreement, we calculated the standard deviation (*SD*) of ratings across analysts for a given construct and scale for each occupation and the SE_M of these ratings. For both indices, lower values indicate higher agreement, and vice versa.

A summary of these results is shown in Appendix C. The columns labeled "Mean of *Ms*" show the mean of the analyst mean importance and level ratings across the 52 abilities for each occupation.[1] The columns labeled "Median of *SDs*" show the median of the *SDs* associated with each mean importance and level rating across the 52 abilities for each occupation. Finally, the columns labeled "Median of SE_Ms" show the median of the SE_Ms associated with each mean importance and level rating across the 52 abilities for each occupation.

The importance ratings across all occupations had a median *SD* of .53 and a median SE_M of .19. The level ratings across occupations had a median *SD* of .76 and a median SE_M of .27. These results are almost identical to those found in Cycle 1. Overall, while the values are generally greater for the level than they are for the importance, the results indicate that the ratings made by the analysts were reasonably consistent for both scales.

[1] While the mean is not a measure of agreement, it can affect the potential range of the *SD* and SE_M.

Cycle 2 Interrater Reliability: Across Constructs Within Occupations

To examine the interrater reliability of the Cycle 2 ratings we calculated the interclass correlations ICC [3, *k*]; Shrout & Fleiss, 1979) among the analyst's ratings to look at consistency across constructs within occupations. As mentioned previously, this calculation examines the similarity in the rank ordering and relative distance between the abilities on a particular scale within an occupation. Our target level of interrater reliability is that the median *ICC* (3, *k*) be .80 or greater. The value of .80 is judged to be a good rule-of-thumb that has been used previously in the O*NET context (e.g., McCloy, Waugh, & Medsker, April 1998).

The results of these analyses are presented in Appendix D. The data revealed high levels of interrater reliability across the 126 Cycle 2 occupations. Specifically, the mean ICC for importance ratings for the abilities across the occupations was .95 (*SD* = .03). The mean ICC for the level ratings was .94 (*SD* = .03). The reliability for both the importance and level ratings exceeded the target coefficient value of .80. Interrater reliability did not vary greatly across occupations and the mean coefficient for importance ratings was just barely higher than the mean level ratings. Results also indicate that occupations with the lowest reliability coefficients for importance had the lowest values for level ratings. This may be due to the skip pattern which forces a "0" for level if the ability is rated not important. This will be monitored when analyzing the data collected in future cycles.

Cycle 2 Interrater Reliability: Across Occupations Within Constructs

Another effective way to evaluate the reliability of the analyst's ratings is to look at the consistency across occupations within constructs. This type of reliability is the extent to which raters agree about the order of and relative distance among occupations on a particular scale for particular construct. For example, is there consistency across raters in how they differentiate among occupations on the required level of the ability *Oral Comprehension*? To make this evaluation, Shrout and Fleiss' (1979) *ICC*(3, *k*) must be calculated for each construct on each scale (instead of for each occupation on each scale as described above). For example, each of the 52 ability importance scale ratings will have a reliability value. The target level of interrater reliability for this coefficient is that the median *ICC*(3, *k*) across the construct ratings for a particular domain on a particular scale be .80 or greater (e.g., the median reliability across 52 ability level ratings should be at least .80). The value of .80 is judged to be a good rule-of-thumb that has been used in the O*NET context before (e.g., McCloy, Waugh, & Medsker, April 1998).

This type of reliability was not used to evaluate the raters during the Cycle 1 data collection process because it should not be calculated until analysts have rated a reasonable number of occupations. However, with the completion of Cycle 2, there were a total of 180 occupations. The results of this analysis are presented in Table 3. The values in the columns titled ICC(C,1) reflect the single rater reliabilities, whereas the values in the columns titled ICC(C,8) reflect the reliability for eight raters. The lowest ICC(C,8) reliabilities were found for speech recognition on both importance and level, and time sharing on level. This may be due to low variation in the importance or the required level of these abilities across jobs or disagreement among raters. However, keep in mind that some variation in calculated values is likely to occur by chance. As previously described, the goal was for the ICC(C,8) reliabilities to have a median value across constructs of .80 or greater. Median ICC(C,8) reliabilities for importance and level were .88 and .90, respectively. These results suggest that there was a good level of agreement among the raters with respect to the order and relative distance among occupations on particular constructs for importance and level.

Table 3. Interrater Reliabilities and Standard Errors of Measurement Across Cycle 1 and 2 Occupations.

Ability		Cycle 1 and 2 (*N* = 180)					
		Importance			Level		
		ICC(C,1)	ICC(C,8)	s_E	ICC(C,1)	ICC(C,8)	s_E
1	Oral Comprehension	0.29	0.77	0.18	0.44	0.86	0.21
2	Written Comprehension	0.50	0.89	0.18	0.56	0.91	0.23
3	Oral Expression	0.33	0.80	0.19	0.44	0.86	0.20
4	Written Expression	0.42	0.85	0.20	0.60	0.92	0.24
5	Fluency of Ideas	0.46	0.87	0.23	0.46	0.87	0.34
6	Originality	0.57	0.91	0.20	0.58	0.92	0.28
7	Problem Sensitivity	0.31	0.78	0.20	0.47	0.87	0.26
8	Deductive Reasoning	0.30	0.78	0.20	0.50	0.89	0.24
9	Inductive Reasoning	0.39	0.84	0.20	0.52	0.90	0.26
10	Information Ordering	0.22	0.69	0.20	0.40	0.84	0.24
11	Category Flexibility	0.26	0.73	0.22	0.30	0.78	0.29
12	Mathematical Reasoning	0.48	0.88	0.24	0.60	0.92	0.33
13	Number Facility	0.43	0.86	0.24	0.53	0.90	0.36
14	Memorization	0.13	0.55	0.23	0.20	0.67	0.36
15	Speed of Closure	0.25	0.73	0.27	0.29	0.76	0.43
16	Flexibility of Closure	0.35	0.81	0.26	0.34	0.81	0.40
17	Perceptual Speed	0.28	0.76	0.28	0.27	0.75	0.38
18	Spatial Orientation	0.52	0.90	0.22	0.50	0.89	0.31
19	Visualization	0.50	0.89	0.25	0.52	0.90	0.39
20	Selective Attention	0.16	0.60	0.21	0.20	0.66	0.26
21	Time Sharing	0.24	0.72	0.21	0.16	0.61	0.30
22	Arm-Hand Steadiness	0.67	0.94	0.22	0.65	0.94	0.30
23	Manual Dexterity	0.62	0.93	0.22	0.55	0.91	0.35
24	Finger Dexterity	0.44	0.86	0.25	0.45	0.87	0.34
25	Control Precision	0.64	0.93	0.21	0.59	0.92	0.34
26	Multilimb Coordination	0.63	0.93	0.22	0.60	0.92	0.31
27	Response Orientation	0.63	0.93	0.19	0.63	0.93	0.31
28	Rate Control	0.65	0.94	0.16	0.67	0.94	0.23
29	Reaction Time	0.67	0.94	0.20	0.67	0.94	0.34
30	Wrist-Finger Speed	0.32	0.79	0.24	0.31	0.78	0.44
31	Speed of Limb Movement	0.53	0.90	0.19	0.52	0.90	0.29
32	Static Strength	0.72	0.95	0.19	0.74	0.96	0.29
33	Explosive Strength	0.49	0.88	0.14	0.54	0.90	0.22
34	Dynamic Strength	0.62	0.93	0.18	0.63	0.93	0.28

Continued on next page

Table 3. Interrater Reliabilities and Standard Errors of Measurement Across Cycle 1 and 2 Occupations.							
Ability		Cycle 1 and 2 (*N* = 180)					
		Importance			Level		
		ICC(C,1)	ICC(C,8)	s_E	ICC(C,1)	ICC(C,8)	s_E
35	Trunk Strength	0.59	0.92	0.22	0.61	0.93	0.27
36	Stamina	0.63	0.93	0.20	0.63	0.93	0.27
37	Extent Flexibility	0.70	0.95	0.19	0.72	0.95	0.33
38	Dynamic Flexibility	0.21	0.68	0.12	0.21	0.69	0.19
39	Gross Body Coordination	0.62	0.93	0.20	0.63	0.93	0.26
40	Gross Body Equilibrium	0.65	0.94	0.16	0.62	0.93	0.24
41	Near Vision	0.19	0.66	0.20	0.40	0.84	0.25
42	Far Vision	0.45	0.87	0.25	0.39	0.84	0.38
43	Visual Color Discrimination	0.47	0.88	0.24	0.51	0.89	0.36
44	Night Vision	0.65	0.94	0.13	0.55	0.91	0.25
45	Peripheral Vision	0.65	0.94	0.15	0.62	0.93	0.22
46	Depth Perception	0.59	0.92	0.21	0.59	0.92	0.29
47	Glare Sensitivity	0.70	0.95	0.13	0.71	0.95	0.21
48	Hearing Sensitivity	0.49	0.89	0.23	0.51	0.89	0.33
49	Auditory Attention	0.39	0.83	0.21	0.41	0.85	0.34
50	Sound Localization	0.54	0.90	0.15	0.59	0.92	0.23
51	Speech Recognition	0.10	0.46	0.25	0.18	0.64	0.33
52	Speech Clarity	0.18	0.64	0.22	0.26	0.74	0.28

Note. These ICCs indicate how consistently raters rated occupations on a given ability.
s_E = Standard error of measurment = Observed score variance times the square root of one minus ICC(C,8).

Summary

The main findings of the analysis of Cycle 2 analyst ratings were as follows:

- The not-relevance and suppression criteria did not generate any results reflecting poorly on the quality of the Cycle 2 ratings.
- While interrater agreement was higher for importance than for level ratings, overall results indicate that the ratings made by the analysts were consistent for both scales across occupations.
- All within-occupation ICC reliabilities were well above the target value of .80 (McCloy, Waugh, & Medsker, April 1998). These high levels of interrater reliability indicate that the analysts rank ordered the abilities within each occupation similarly on both importance and level.
- Index interrater reliability calculated at the end of Cycle 2 did not vary greatly from one occupation to the next.
- The importance and level median across-occupation ICC reliabilities were above the target value of .80. These high levels of interrater reliability indicate that analysts rank ordered occupations within each ability similarly on both importance and level.

Given these results, it appears as though the analysts were well trained and generally understand the abilities and associated definitions. Agreement was high and there is clear evidence regarding the quality of the data. Nevertheless, it may be beneficial to closely examine the definitions, as well as the training, associated with the abilities that were flagged more often than others for having a $SE_M > .51$ for level (e.g., wrist-finger speed, far vision, flexibility of closure). It's possible that additional clarification could reduce the observed variance for those abilities.

References

Donsbach, J., Tsacoumis, S., Sager, C., & Updegraff, J. (2003). *O*NET analyst occupational abilities ratings: Procedures* (DFR-03-22). Alexandria, VA: Human Resources Research Organization.

Fleishman, E.A., Costanza, D. P, & Marshall-Mies, J. (1999). Abilities. 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* (p.175-195). Washington D.C.: American Psychological Association.

McCloy, R., Waugh, G., & Medsker, G. (1998, April). *Determining the occupational reinforcer patterns for O*NET occupational units*. Alexandria, VA: Human Resources Research Organization.

Noble, C.L., Sager, C., Tsacoumis, S., Updegraff, J. & Donsbach, J. (12003).*O*NET analyst occupational abilities ratings: Wave 1 results.* Alexandria, VA: Human Resources Research Organization.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428