Factors of Numeracy and Literacy Proficiency for South Korean Senior Workers

Zhen Liu, Yutong Liu, Jingyi Huang

March 23, 2018

1 Summary

Our study aims to identify how demographic, organizational and learning factors are associated with numeracy and literacy scores of senior workers in South Korea. After applying model selection methods, the result shows that gender, education level, work flexibility, active learning strategies, public or private sectors are important variables associated with both numeracy scores and literacy scores.

2 Introduction

The senior population in South Korea is rapidly growing. As a result, seniors continue to participate in the labour market, and often engage in educational programs to further develop their workplace skills. It would be interesting to identify what is the association between different factors of senior workers and their skills in general. However, not all factors are equally useful. Demographic and organizational factors are intrinsic characteristics of the senior workers, which are usually hard to change. Thus, it is not possible to improve senior workers' skills by changing those factors. In contrast, learning variables (eg. Participation in job-related education, Participation in non-job-related education) are easy to change. People can be encouraged to attend the educational program. Knowing which learning factors are associated with senior workers' skills will help Human Resource officers to develop such training programs and further improve workers' literacy and numeracy skills.

From the data analysis, we expect meaningful findings to apply to the development of senior workers. At the same time, we expect our study findings contribute the field of adult education and organization development.

In the next part of the report, we include the detailed description of the data, methods used to select variables and the corresponding results. In the end, we make interpretations about the final model selected.

3 Data description and Explanatory analysis

The data set was obtained from the open source of the Programme for the International Assessment of Adult Competencies (PIAAC). The data were collected from 24 countries between August 2011 and March 2012, and each country drew a representative sample of individuals aged from 16 to 65 years old. Based on the objective of the study, only employees aged from 50 to 65 years old in South Korea were included. As a result, 1247 data entries were used for the initial analysis.

3.1 Response(Dependent) variables

Two response variables are the test score of numeracy proficiency(numeracy score) and the test score of literacy proficiency(literacy score). Histograms in Figure 1 provide a good visualization of the distribution of each response variable. Numeracy score ranges from 105 points to 381 points, whereas literacy score ranges from 119 points to 372 points.

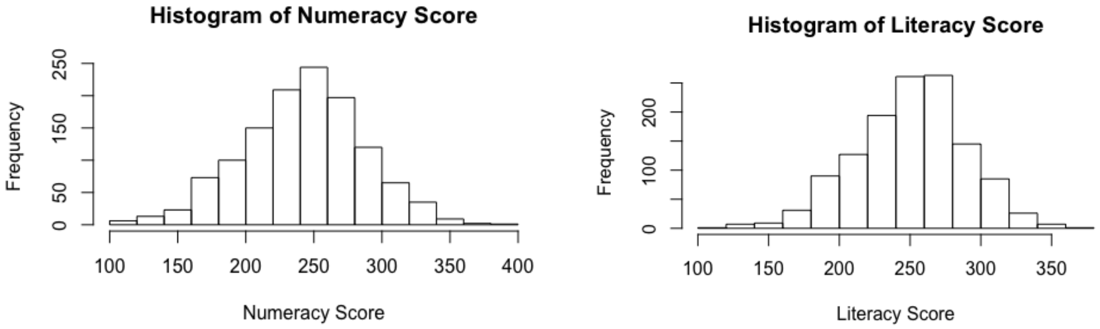


Figure 1: Histograms of numeracy and literacy test scores

3.2 Explanatory(Independent) variables

There are 17 independent variables, which belong to three different categories: demographic, organizational and learning factors. Table 1 lists all variables with abbreviations for each category.

Demographic Factors	**Organizational Factors**	**Learning factors**
Continuous: Age Work Experience in years(WorkY)	*Continuous*: Work Flexibility(Flex) Learning opportunity(Oppo)	*Continuous*: Active learning strategies(Active) Hours of participation in non-formal education(Hour)
Categorical: Gender Education level (EDLevel) Employment type: Full-time or part-time(Full)	*Categorical*: Private/ public sector (Priv) Manage others (Mgr) Numer of Managing employees (Mgr_c)	*Categorical*: Participation in: non-formal education (NFE) adult education (AE) job-related education (EJR) job-related adult education(AEJR) non-job-related adult education (AENJR)

Table 1: Description of independent variables

Work flexibility, learning opportunity and active learning strategies are the average of likert scales. Overall, higher work flexibility value means that the employee has more freedom to adjust their work schedule and work hour. Learning opportunity captures how likely an employee is able to learn from the work. If a worker has a high value for learning opportunity, it means that he or she has good opportunity to learn new knowledge when doing the job. Active learning strategies captures an employee's ability to master the learning strategies.

3.3 Missing values

There are 501 missing values in the *Mgr* column, which is more than 40 percent of the total number of observations. Besides, there are only about 100 observations that have values for *Mgrc* column. Due to large amount of missing values, neither of these two columns are very informative. Thus, we removed them from the dataset. Moreover, there are 39 missing values in the *Priv* column, which is about 3 percent of the total observations. Since it is relatively small compared to the whole dataset, it makes sense to remove those observations rather than the whole column. After deleting the observations with unknown sectors, 1207 observations were used for the final analysis.

3.4 Data pre-processing

Table 2 shows that for *EDLevel*, there are only one observation in the research category and four observations in the post-graduate category. To solve this unbalanced data problem, we merge college, post-graduate and research into one category: college or higher. Thus, EDLvel has three levels instead of five in the end.

Category	Middle school	High school	College	Post-graduate	Research
observations	501	440	261	4	1

Table 2: Summary statistics of Education level

We also noticed that AE is the union of the other two independent variables: AEJR and AENJR. Having multi-collinear variables will lead to inaccurate results of variable selections. After conducting statistical tests, results show that AE, AEJR and AENJR all have effects on the test scores. Thus, we removed AE from the model. Also, we checked that AEJR and EJR have exact same values. In order to avoid extremely high correlation, we removed EJR.

In the end, there are 13 variables left for the final analysis. Summary Statistics for categorical variables is shown in the Table 3.

Variables	**Categories**	**Size**	**Mean for literacy**	**Mean for numeracy**
Gender	1: Male	756	255.2	248.0
	2: Female	468	241.8	231.3
Education Level	1: Middle school	501	230.0	217.0
	2: High school	440	254.8	246.6
	3: College or higher	266	280.4	280.3
Employment Type	1: Full-time	1027	251.8	243.5
	2: Part-time	180	240.7	231.7
Private/Public Sector	1: Private	1062	247.5	238.4
	2: Public	145	269.3	266.6
Non-formal Education	0: Participation	694	242.1	231.1
	1: No	513	261.0	256.1
Job-related Adult Education	0: Participation	801	244.9	234.9
	1: No	406	260.5	255.3
Non-job-related Adult Education	0: Participation	1097	234.9	239.9
	1: No	110	255.3	260.0

Table 3: Summary Statistics: Categorical Variables

3.5 Boxplots

We use boxplots to visualize the data. Since there are too many variables in our dataset and putting all of them in the report will be overwhelming for our client, we choose education level as an example to interpret the boxplots. According to the boxplots in Figure 2, the median of both literacy and numeracy score increases as the education level increases. This indicates a positive association between the education level and the two scores.

3.6 Correlation

We checked the correlations within categorical variables and continuous variables respectively. Results in Figure 3b show that AEJR and NFE have very strong correlation (r=0.8) The literacy score and numeracy score are also highly correlated (r=0.9). Besides, in Figure 3a, the bigger and darker the circles are, the higher the correlation between the two variables. For example, the circles between active learning and the two scores are large which indicates that active learning is a potential important variable that are highly associated with the scores.

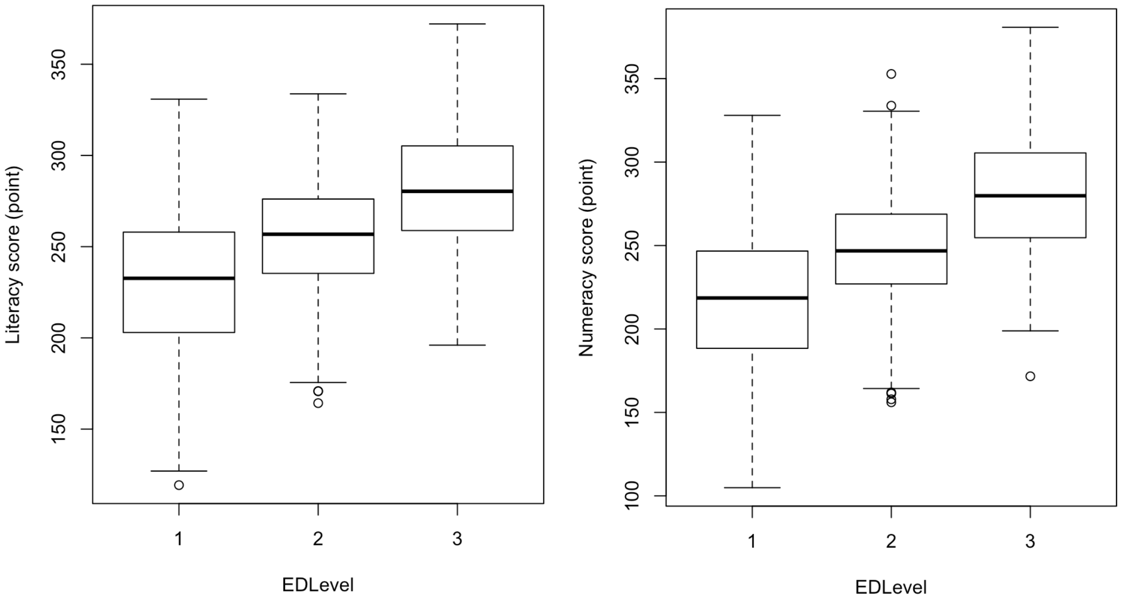
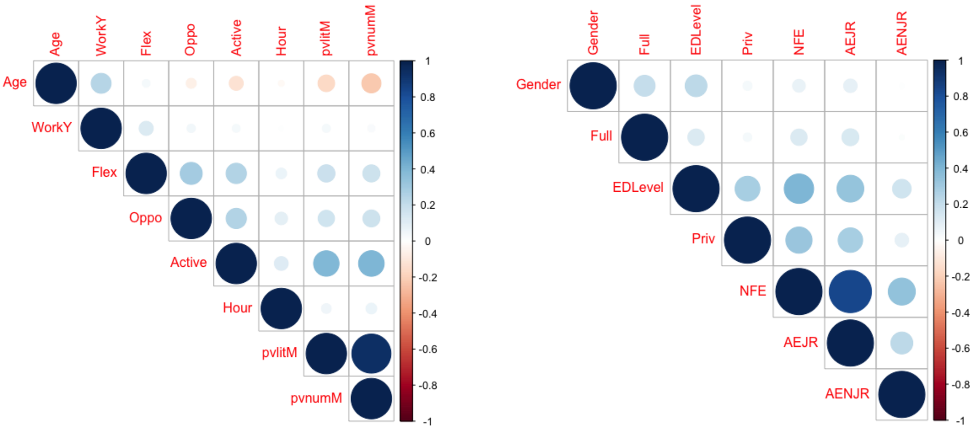


Figure 2: Boxplots of EDLevel vs literacy and numeracy scores.



(a) Continuous Variables

(b) Categorical Variables.

Figure 3: Correlation plots for continuous variables and categorical variables

4 Method

4.1 Stepwise AIC

Stepwise AIC is one of the most popular variable selection methods. AIC is the criterion used to compare two models. AIC increases as the goodness-of-fit improves but decreases as the number of parameters increases. Therefore, it helps to find a model that is both simple and fits the data well. Stepwise is the algorithm to search for candidate models. The naive approach is to search all possible subsets of the variables, but it is time-consuming and sometimes infeasible. Stepwise method helps to solve this problem. We used sequential replacement for the analysis, which is one type of stepwise methods. Essentially, sequential replacement starts from a model with no variables, and adds one variable that improves the criterion most significantly to the model. As it keeps adding variables to the model, it also considers removing variables from the model that become insignificant. The process keeps going until the performance under a certain criterion does not improve anymore.

4.2 Nested model comparison: ANOVA

ANOVA can be used to compare two nested models. For example, when comparing model 1 (A+B+C+D) and model 2 (A+B+C), we use ANOVA to perform F test. It tests whether adding factor D is significant or not. Small p-value of the test indicates that factor D can provide important information. On the other hand, if the p-value is not small enough, these two models do not have a big difference. In this case, model 2 would be a better choice due to parsimony.

5 Result

5.1 Variable Selection for Literacy Score

Model I in Table 4 is the best model that the sequential selection method picked because it has the smallest AIC. However, the estimated coefficient of variable $Hour$ in model I is -0.003 (close to zero). This suggests that by holding other factors constant, if $Hour$ increases by 1000 units, then the test score will decrease by 3. It means that a large change in $Hour$ will lead to little change in test scores. Also, the p-value of estimated coefficient of variable $Hour$ is 0.11, which is not significant. Due to the above reasons, it might be better to exclude $Hour$.(See Model II in Table 4)

Model	Variables	Adjsted R^2	AIC
I	Gender + Age + EDLevel + Flex + Priv + Active + Hour	0.304	8402.8
II	Gender + Age + EDLevel + Flex + Priv + Active	0.303	8403.4

Table 4: Model Comparison for Literacy Score

To verify this, we use ANOVA to Compare Model I and II. The result shows that the p-value is 0.12. This suggests that these two models do not have a big difference; in other words, the additional variable (Hour) in Model I does not provide extra information than model II. Also, their adjusted R^2 and AIC have few differences. Therefore, for parsimony, we would recommend Model II as the best model.

5.2 Variable Selection for Numeracy Score

For numeracy score, Model I in Table 5 has the smallest AIC throughout the Sequential selection process. Same as for literacy score, the p-value of ANOVA Comparison between model I and II is 0.17. We suggest Model II as the best model due to the same reason in Section 5.1.

Model	Variables	Adjsted R^2	AIC
I	Gender + Age + EDLevel + Flex + Priv + Active + Hour	0.3564	8626.0
II	Gender + Age + EDLevel + Flex + Priv + Active	0.3557	8626.3

Table 5: Model Comparison for Numeracy Score

5.3 Coefficients Interpretation

After fitting a linear regression model with all the important variables picked by the previous methods: education level, active learning, work flexibility, age, gender and public/private sector, we get Table 6. All the variables are significant at 0.05 significance level ($p<0.05$). Education levels have positive estimated coefficients indicating that the higher the education level the senior workers have, the higher the scores they receive. For example, we observe that the workers with college school education receive higher scores than the workers with middle school education if we hold other variables constant. This result agrees with the increasing trend of median in Figure 2. Similarly, given positive estimated coefficients, the higher ability of mastering active learning strategies and the more flexible they arrange their tasks, the better the scores they get. Besides, workers from the public sector tend to have higher scores than the workers in private sector from our data. Given negative estimated coefficients, the scores decreases with the increase of the worker's age, and women receive lower scores compared with men if we hold other variables constant.

Variable	Literacy Score		Numeracy Score	
	Estimated Coefficient	p-value	Estimated Coefficient	p-value
Intercept	238.57	<0.01	259.60	<0.01
High vs Middle School	19.43	<0.01	22.52	<0.01
College vs Middle School	36.43	<0.01	46.77	<0.01
Active learning	8.49	<0.01	8.72	<0.01
Work Flexibility	1.92	0.010	1.90	0.020
Age	-0.57	0.011	-1.16	<0.01
Female vs Male	-4.07	0.041	-5.64	<0.010
Public vs Private	6.68	0.027	9.25	<0.01

Table 6: Estimated Coefficients for Independent variables

6 Conclusion and Discussion

In conclusion, gender, age, education level, work flexibility, public/private sector and active learning strategies are highly associated with both proficiency scores(numeracy and literacy). The correlation between the numeracy score and the literacy score is 0.93. This explains why the same variables are picked for both scores.

In practice, flexibility in the workplace could help senior employees improve their proficiency of numeracy and literacy skills. Employees who have higher education level are more likely to get higher scores in the proficiency tests. To improve senior workers' skills, the Human Resource department can focus on developing training programs about active learning strategies.

Balanced data usually provide better estimation of the coefficients. From the summary statistics (Table 3), the size of each category for some variables is much different. This unbalancedness may affect the significance of the analysis since a much larger weight is given to one of the category. Moreover, in the *Priv* variable, there are 39 missing values and we excluded them, which may introduce bias in the results. The performance of our result might be better if we could use some algorithm to do data imputation, such as K-NN(using the category of its neighbor to classify it).

A Appendices

A.1 Appendix 1

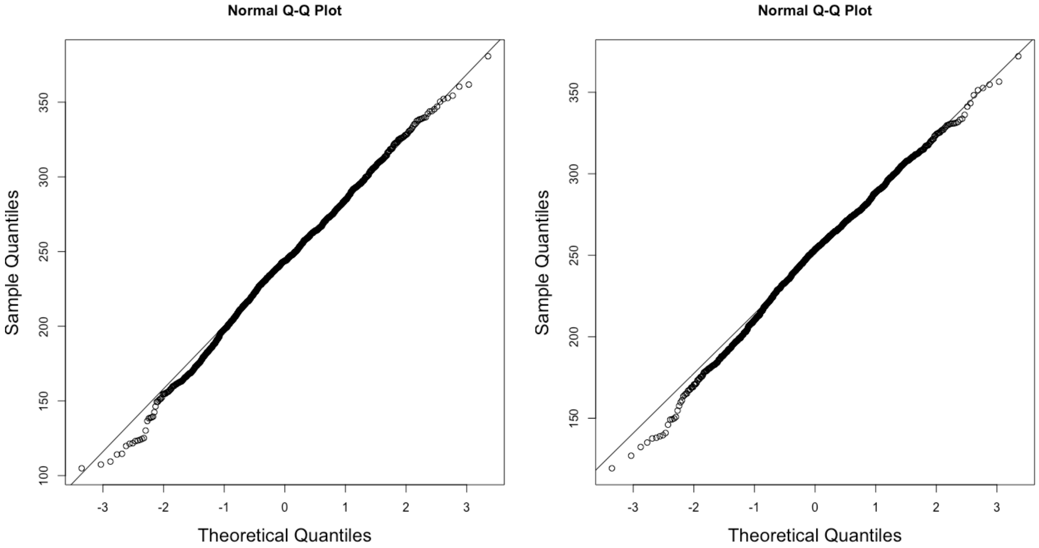


Figure 4: Q-Q Plots for dependent variables

From Figure 4, two Q-Q plots both seem on a straight line so it is fair to say that the two responses(Literacy score, Numeracy score) of our sample are normal distributed. The assumption of linear regression is satisfied.

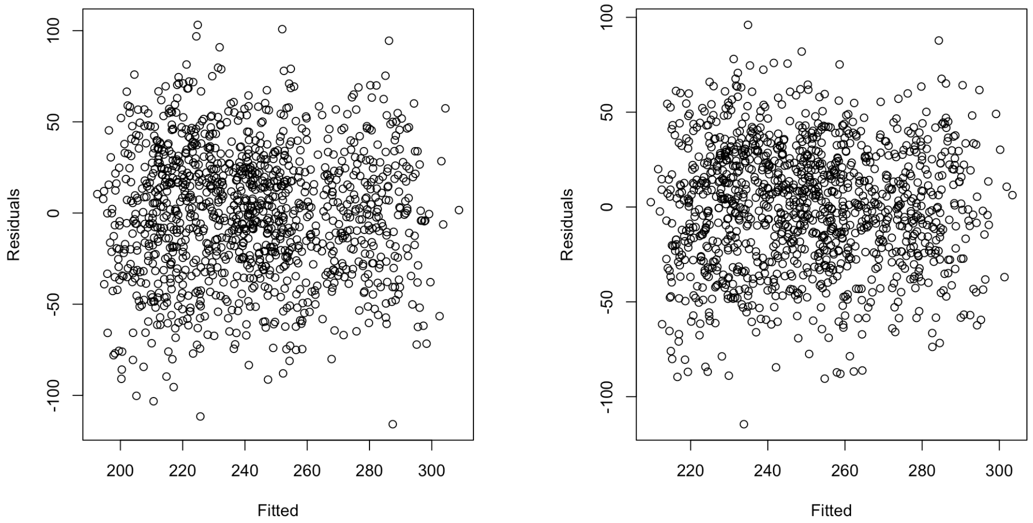


Figure 5: Residual plots for numeracy score (left) and literacy score (right)

The data points in the residual plot in Figure 5 are scattered around zero. Hence, the equal variance assumption is not violated.

A.2 Appendix 2: Sequential Selection-Adjusted R^2

Using multiple approach to do model selection could provide further evidence supporting the inclusion or exclusion of certain variables. We tried an another criterion - Adjusted R^2- in the variable selection section. It can seek a model that has a good fit to the truth and with relatively few parameters. It increases only if the new variable added improves the model more than would be expected by chance.

For literacy score, the model selected under Adjusted R^2 is exactly the same as that under AIC criteria (See Table 4, Model I).

For Numeracy score, Model I in Table 7 is preferred. Its p-values of Hour, NFR, AEJR, AENJR and WorkY are all larger than 0.05, which state that employees under different levels of these variables probably have same test results.

Model	#	**Variables**	**Adjsted R^2**	**AIC**
I	11	Gender + Age + EDLevel + Flex + Priv + Active + Hour + NFE + AEJR + AENJR + WorkY	0.3577	8626.2
II	6	Gender + Age + EDLevel + Flex + Priv + Active	0.3557	8626.3

Table 7: Model Comparison for Numeracy Score

Nested ANOVA Comparison is processed. Table 8 shows the ANOVA Comparison between model II and other models in order to figure out if Hour, NFR, AEJR, AENJR or WorkY are needed. All the p-values are above 0.05, which suggests that these models do not have big differences and not provide additional effective information compared to model III. Also, their adjusted R^2 and AIC have slightly differences. Therefore, we would recommend Model II as the 'best' model for parsimony.

Model	**Variables**	P-value
II	Gender + Age + EDLevel + Flex + Priv + Active	-
I	Gender + Age + EDLevel + Flex + Priv + Active + Hour + NFE + AEJR + AENJR + WorkY	0.16
III	Gender + Age + EDLevel + Flex + Priv + Active + Hour	0.17
IV	Gender + Age + EDLevel + Flex + Priv + Active + NFE	0.46
V	Gender + Age + EDLevel + Flex + Priv + Active + AEJR	0.75
VI	Gender + Age + EDLevel + Flex + Priv + Active + AENJR	0.31
VII	Gender + Age + EDLevel + Flex + Priv + Active + WorkY	0.40

Table 8: ANOVA Comparison for Numeracy Score