Analysis of Mathematics Scores in Achievement Exam of Information Technology College Students Using Covariance Structure Analysis

A. Ishida*, N. Yamamoto, and J. Murakami

***Abstract*—In recent years, big data has been attracting attention in various fields, and Japanese institutions of higher education are also focusing on improving education based on the analysis of data accumulated in schools. Therefore, as an example of the application of big data analysis to mathematics education, we considered the mathematics exam score data accumulated at our college as big data and analyzed it. Another major objective is to obtain some knowledges about mathematics education for information engineering students. The software used for the analysis is the freeware statistical analysis software R. The analysis method used is covariance structure analysis, which represents correlations and causal relationships among observed variables as a well-fitting model. As a result, two models with good fit indices were created for the integrated data of multi-year exams, and it was found the learning items that have a significant impact on the later learning items in that data. Furthermore, it was also able to use one of the models to analyze the characteristics of a single-year data. It is thought that these results can contribute to mathematics education for information engineering course students.**

***Index Terms*—Covariance structure analysis, achievement exam, mathematics education, R, information course students**

I. INTRODUCTION

In recent years, huge amount of information has been gathered in various industries and situations through websites and the IoT equipment. Accumulated data has come to be used as big data, and it is important to share its usage and analysis methods [1]. We have been studying various analysis methods for big data and analyzing rehabilitation data and sleep data using these methods [2, 3]. We have also considered big data as a tensor (multidimensional array) and developed educational materials to teach tensor decomposition methods [4–7].

Educational improvement based on the analysis of data accumulated in schools has been attracting attention as "Institutional Research (IR)" in Japanese higher education institutions [8]. Therefore, we focused on the learning data (examination score data) accumulated at our schools and conducted an analysis. Mathematics and science examinations called "Achievement Exam" had been carried out to the third grades students (17-18 years old) until 2017 at all 51 colleges of the National Institute of Technology, including our college. The purpose of this examination is to improve educational contents and methods, and to encourage students to develop proactive learning attitudes. The contents of this examination correspond to the contents that the Institute have indicated as the minimum goals for students to achieve as part of their educational quality assurance (Model Core Curriculum, MCC) [9]. We thought that the analysis of this mathematics exam score data as big data was expected to be useful for mathematics education for engineering course students, especially information engineering course students, and was an important example of the use of big data.

In this study, we analyze the mathematics achievement exam score data for multiple years accumulated on our campus and investigate the relationship between the learning items of mathematics studied by the second year. Specifically, based on the assumption that there exist specific models that fits well with the learning items regardless of the year, we create models that fits well with the integrated multi-year data, interpret its meaning, and check the goodness of fit of the model in a single year using the goodness-of-fit indices. The analysis is performed using the freeware statistical analysis software R [10] so that similar analyses can be performed widely in general. The analysis method used is covariance structural analysis, also known as SEM, which can represent correlations and causal relationships among observed variables as a well-fitting model [11].

II. EXAMINATION SCORE DATA USED

The data used in this study is the mathematics score data of the learning achievement examination from 2014 to 2017 at National Institute of Technology (NIT), Kumamoto College. This exam is a mark sheet test consisting of the following ten learning items: "Calculation of Numbers and Expressions", "Equations and Inequalities", "Functions and Graphs", "Number of Events and Sequences", "Properties of Plane Vectors", "Calculation of Differentiation and Integration", "Applications of Differentiation and Integration", "Space Vectors and Matrix Properties", "Eigenvalues and Matrix Expressions", and "Differentiation and Integration of Functions of Two Variables". Each item is scored out of 50 points. At each NIT, this exam was performed by selecting items from those items according to the learning progress of each college.

This paper dealt with the score data for the following six items studied by the second grade at our college: "Calculation of Numbers and Expressions" (*Alg*), "Equations and Inequalities" (*Eq*), "Functions and Graphs" (*Fun*),

Manuscript received September 25, 2022; revised October 23, 2022; accepted December 5, 2022.

A. Ishida is with the Faculty of Liberal Arts, Kumamoto College, National Institute of Technology, Koshi, Japan.

N. Yamamoto and J. Murakami are with the Faculty of Electronics and Information Systems Engineering, Kumamoto College, National Institute of Technology, Koshi, Japan.

*Correspondence: shida@kumamoto-nct.ac.jp (A.I.)

doi: 10.18178/ijiet.2023.13.3.1821

"Number of Events and Sequences" (*Seq*), "Properties of Plane Vectors" (*Vec*), and "Calculation of Differentiation and Integration" (*Cal_bas*). Note that the symbols in parentheses correspond to variable names.

Fig. 1 shows the data structure used in this analysis. As shown in this figure, the score data is handled as a matrix (two-dimensional array), with learning items in columns and individual students in rows. Data for each fiscal year from 2014 to 2017 and integrated data for all of them were used for the analysis data.

Table I and Fig. 2 show basic statistics and box plots of total scores for the six learning items for each data set. These statistics show that the mean score is high for students in 2014, and the sample standard deviation in 2015 is relatively large, indicating a large variability in understanding among students in this year. Fig. 2 also shows that there are more outliers in 2014 than in other years.

	Alg	Eq	Fun	Seq	Vec	Cal_bas
Student 1						
Student 2						
⋮						

Fig. 1. Structure of score data for analysis.

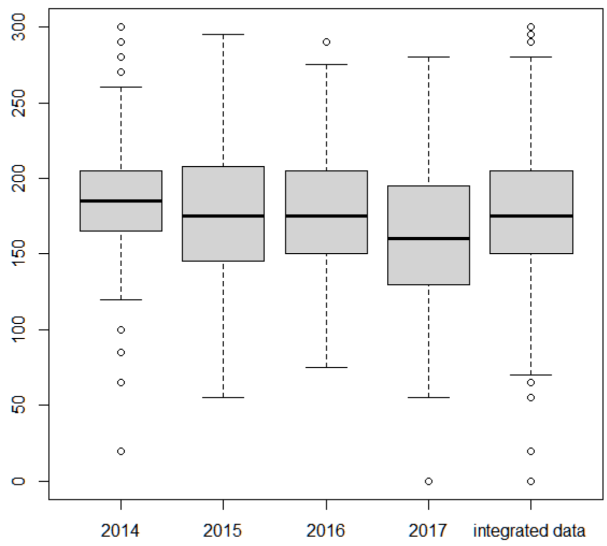

Fig. 2. Box plots of each data.

TABLE I: BASIC STATISTICS

	Number of students	mean value	median	sample standard deviation
2014	116	184.7	185	42.9
2015	131	175.3	175	49.1
2016	131	177.6	175	41.7
2017	126	164.3	160	47.2
integrated data	504	175.3	175	45.9

Fig. 3 shows a heatmap of the correlation matrix obtained from the integrated data using R described in next chapter. From this figure, the correlations between *Fun* and *Cal_bas* variables and between *Seq* and *Vec* variables are particularly high.

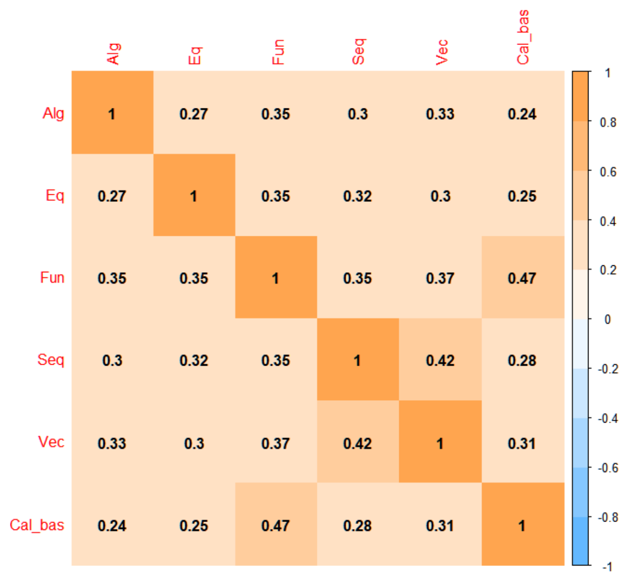

Fig. 3. Heat map of the correlation matrix.

In addition, the normality of the integrated data was confirmed using "shapiro.test" which is the function of R for the Shapiro-Wilk test. As a result, the *p*-value was 0.06102, and the null hypothesis of normal distribution could not be rejected.

III. Statistical Analysis Software R and Covariance Structure Analysis

A. Statistical Analysis Software R

As mentioned in Chap. 1, the statistical analysis software R is freeware, and its use has been spreading along with the recent focus on data science. Various programs and materials can be obtained from archive called CRAN and its mirror sites via the Internet. It is possible to use them to perform many statistical processes and to create figures. This software has feature to handle vectors, matrices, and multidimensional arrays, and the ability to exchange data with Microsoft Excel makes it an easy-to-use software for statistical processing.

B. Covariance Structure Analysis

To analyze the relationships among the items in the score data, we used a method called covariance structure analysis, also called structural equation modeling (SEM). This is a method of analyzing original data by creating a model that represents causal relationships only among observed variables or causal relationships assuming latent variables as arrows with path coefficients and interpreting the meaning of the model. This method is used in various field including educational research [12, 13].

In this study, under the assumptions described in Chap. 1, we used the learning items listed in Chap. 2 as observed variables for the integrated data. Then, we created models that well represents the causal relationships among them and analyzed the data. The results of the analysis are showed in the next chapter. The "sem" function provided the sem package of R was used for this analysis [14], and the path diagram was created with the "pathDiagram" function of DiagrammeR package [15].

IV. Modeling and Discussion of Integrated Data

In this chapter, we create a model that fits the integrated data well. The following hypotheses were made in creating the model.

1) The comprehension of previously learned contents influences that of newly learned contents.
2) The paths exist between the learning items with high correlation coefficients.

The following two models were created, referring to the above hypotheses and the correlation coefficients between the variables shown in Fig. 3. We named them model_1 and model_2, and the structural equations of each model are shown below:

[model_1]

$$\begin{cases} Eq = a_{11} \times Alg + e_1 \\ Fun = a_{21} \times Alg + a_{22} \times Eq + e_2 \\ Seq = a_{31} \times Alg + a_{32} \times Eq + a_{33} \times Fun + e_3 \\ Vec = a_{41} \times Alg + a_{42} \times Eq \\ \qquad + a_{43} \times Fun + a_{44} \times Seq + e_4 \\ Cal\_bas = a_{52} \times Eq + a_{53} \times Fun + a_{54} \times Seq + e_5 \end{cases} \tag{1}$$

[model_2]

$$\begin{cases} Eq = a_{11} \times Alg + e_1 \\ Fun = a_{21} \times Alg + a_{22} \times Eq + e_2 \\ Seq = a_{31} \times Alg + a_{32} \times Eq + a_{33} \times Fun + e_3 \\ Vec = a_{41} \times Alg + a_{42} \times Eq \\ \qquad + a_{43} \times Fun + a_{44} \times Seq + e_4 \\ Cal\_bas = a_{53} \times Fun + a_{54} \times Seq + e_5 \end{cases} \tag{2}$$

where each a_{ij} represents the path coefficient, and each e_k denotes the error. A script example that analyzes model_1 using the "sem" function and and creates a path diagram using the "pathDiagram" function is as follows.

[(Script 1) SEM and creation of path diagram]

```
# Load sem library (only at the first time)
library( sem )
# Load DiagrammeR library (only at the first time)
library( DiagrammeR )
dat <- read.csv( "score.csv" )   # Read data
r <- cor( dat )   # Compute correlation matrix
# Extract number of rows (number of students)
n <- nrow( dat )
# Create a model 1
model_1 <- specifyEquations()
Eq = a11*Alg
Fun = a21*Alg + a22*Eq
Seq = a31*Alg + a32*Eq + a33*Fun
Vec = a41*Alg + a42*Eq + a43*Fun + a44*Seq
Cal_bas = a52*Eq + a53*Fun +a54*Seq
V( Cal_bas ) = e1
V( Vec ) = e2
C( Vec, Cal_bas ) = e3
# Apply model
ans <- sem( model_1, r, N=n, fixed.x="Alg" )
# Display analysis results
summary( ans,rsquare=T, fit.indices=c( "GFI", "AGFI",
"SRMR", "RMSEA", "AIC", "BIC" ) )
# Draw a path diagram
pathDiagram(ans,min.rank="Alg", max.rank="Vec,Cal_bas",
same.rank="Eq,Fun", ignore.double=FALSE,
edge.labels="values", digits=3, node.font=c( "osaka", 10 ) )
```

(End of Script)

A similar script was executed for model_2, and the obtained goodness-of-fit indices for each model are shown in Table II. Regarding the indices, chi-square tests were performed under the null hypothesis that "the constructed model is correct". The *p*-values of model_1 and model_2 calculated from the chi-square values and degrees of freedom *df* are 0.118 and 0.111, respectively. Therefore, the *p*-values of both models cannot reject the null hypothesis at the significance level of 5%, indicating that "the models cannot be said to be incorrect."

Now we consider the indices of two models shown in Table II. GFI (Goodness of Fit Index) and AGFI (Adjusted Goodness of Fit Index) take values between 0 and 1, and a good model is judged to be greater than 0.95. The GFI values of model_1 and model_2 are 0.999 and 0.997, respectively. The AGFI values of them are 0.976 and 0.970, severally. RMSEA (Root Mean Square Error of Approximation) and SRMR (Standardized Root Mean Square Residual) are the better the goodness of fits when the former is less than 0.05 and the latter is closer to 0. The RMSEA values of both models are 0.0381 and 0.0488, respectively, and the SRMR values of them are 0.0103 and 0.0176, severally. Thus, from the above, the created models are considered to represent the original data well.

TABLE II: Goodness-of-Fit Indices

	df	*p*-value	GFI	AGFI	RMSEA	SRMR
model_1	1	0.188	0.999	0.976	0.0381	0.0103
model_2	2	0.111	0.997	0.970	0.0488	0.0176

Table III and Fig. 4 show the path coefficients and path diagram of model_1. From the calculation results, we consider the degree of influence of each item on the "Calculation of Differentiation and Integration" (*Cal_bas*). The coefficients for *Cal_bas* representing the direct effect from *Eq*, *Fun*, and *Seq* are 0.070, 0.406, and 0.121, respectively. Furthermore, since the indirect effect from *Eq* via *Fun* and *Seq* is $a_{22} \times a_{53} + a_{22} \times a_{33} \times a_{54} + a_{32} \times a_{54} \approx 0.143$, the coefficient indicating the total effect from *Eq* is 0.070+0.143=0.213. Similarly, since there is an indirect effect of $a_{33} \times a_{54} \approx 0.026$ from *Fun* via *Seq*, the coefficient showing the total effect form *Fun* is 0.406+0.026=0.432. From *Alg*, there is only an indirect effect, and its coefficient is 0.198. Therefore, *Cal_bas* is most influenced by *Fun*. Specifically, when the score of *Fun* increases by 1 point, the score of *Cal_bus* increases by 0.432 points.

Equally, the coefficients representing the total effect from *Alg*, *Eq*, *Fun*, and *Seq* on the "Properties of Plane Vectors" (*Vec*) are 0.328, 0.223, 0.244, and 0.283, respectively. From this result, *Vec* is most affected by *Alg*. In fact, when *Alg* increases by 1 point, *Vec* increases by 0.328 points.

TABLE III: PATH COEFFICIENTS OF MODEL_1

a_{11}	a_{21}	a_{22}	a_{31}	a_{32}	a_{33}	a_{41}
0.272	0.275	0.274	0.171	0.200	0.216	0.152
a_{42}	a_{43}	a_{44}	a_{52}	a_{53}	a_{54}	
0.099	0.183	0.283	0.070	0.406	0.121	

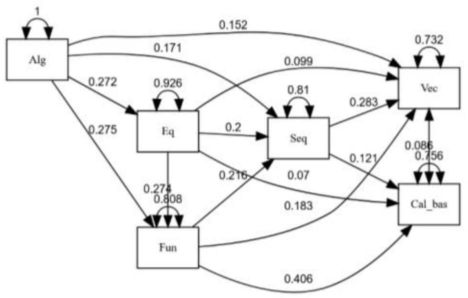

Fig. 4. Path diagram of model_1.

Table IV and Fig. 5 show the path coefficients and path diagram for model_2. A similar analysis using this model shows that *Cal_bas* is particularly influenced by *Fun*. It was found that when *Fun* increased by 1 point, *Cal_bus* increased by 0.432 points. Moreover, *Vec* is the most affected by *Alg*, increasing by 0.327 points when *Alg* increases by 1 point. Also, in model_2, the path coefficient a_{52} of model_1 is removed. Since this path is from *Eq* to *Cal_bas*, model_2 reduces the value representing the total effect from *Eq* to *Cal_bas* to 0.152.

TABLE IV: PATH COEFFICIENTS OF MODEL_2

a_{11}	a_{21}	a_{22}	a_{31}	a_{32}	a_{33}	a_{41}
0.272	0.275	0.274	0.171	0.200	0.216	0.152
a_{42}	a_{43}	a_{44}	a_{53}	a_{54}		
0.091	0.185	0.285	0.425	0.137		

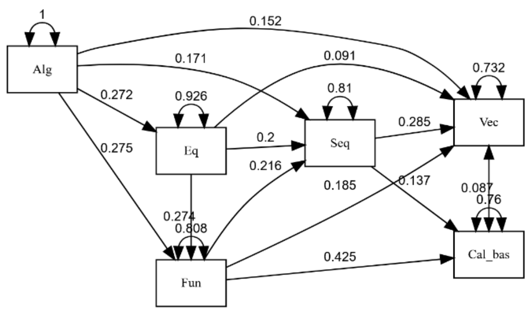

Fig. 5. Path diagram of model_2.

V. APPLICATION OF THE MODEL TO SINGLE YEAR DATA

The model_1 and model_2 created in Chap. 4 are applied to the score data from 2014 to 2017 using Script 1 described in the previous chapter to create the path diagrams.

Tables V and VI show the goodness of fit indices. The values of the indices shown in the tables indicate that both model_1 and model_2 represent the 2014 and 2016 data relatively well. The *p*-value of model_2 is 0.0596 for the 2017 data, and the null hypothesis cannot be rejected at the significance level of 5%.

Table VII and Fig. 6 show the path coefficients and path diagram when model_1 is applied to the 2014 data, which seems to be the best fit based on the values in Table V. *Cal_bas* is most affected by *Fun*, as in the integrated data, and 1 point increase in *Fun* results in 0.305 points increase. For *Vec*, unlike the results for the integrated data, it is particularly influenced by *Fun*, showing an increase of 0.457 points for 1 point increase in *Fun*.

TABLE V: GOODNESS OF FIT INDICES OF MODEL_1 FOR SINGLE-YEAR DATA

year	*df*	*p*-value	GFI	AGFI	RMSEA	SRMR
2014	1	0.738	1.000	0.993	0.000	0.00574
2015	1	0.0118	0.984	0.672	0.203	0.0363
2016	1	0.268	0.997	0.934	0.0419	0.0170
2017	1	0.0177	0.986	0.696	0.192	0.0374

TABLE VI: GOODNESS OF FIT INDICES OF MODEL_2 FOR SINGLE-YEAR DATA

year	*df*	*p*-value	GFI	AGFI	RMSEA	SRMR
2014	2	0.242	0.992	0.915	0.0603	0.0311
2015	2	0.000737	0.966	0.645	0.219	0.0613
2016	2	0.473	0.996	0.960	0.000	0.0191
2017	2	0.0596	0.986	0.848	0.121	0.0377

TABLE VII: PATH COEFFICIENTS OF MODEL_1 FOR 2014 DATA

a_{11}	a_{21}	a_{22}	a_{31}	a_{32}	a_{33}	a_{41}
0.260	0.358	0.169	0.089	0.271	0.337	-0.061
a_{42}	a_{43}	a_{44}	a_{52}	a_{53}	a_{54}	
0.218	0.347	0.326	0.154	0.281	0.071	

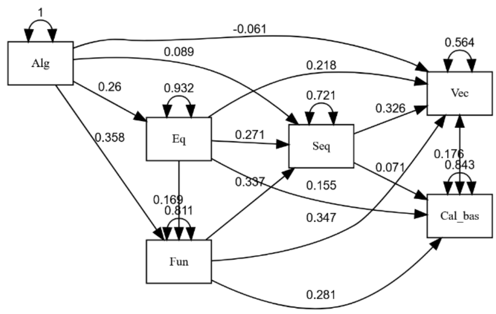

Fig. 6. Path diagram of model_1 for 2014 data.

Here, we created a well-fitting model for the 2015 score data, which did not fit either model well. The structural equations representing the model are shown below:

$$\begin{cases} Eq = a_{11} \times Alg + e_1 \\ Fun = a_{21} \times Alg + a_{22} \times Eq + e_2 \\ Seq = a_{32} \times Eq + a_{33} \times Fun + e_3 \\ Vec = a_{41} \times Alg + a_{42} \times Eq \\ \qquad + a_{43} \times Fun + a_{44} \times Seq + e_4 \\ Cal\_bas = a_{51} \times Alg + a_{52} \times Eq \\ \qquad + a_{53} \times Fun + a_{54} \times Seq + e_5 \end{cases} \tag{3}$$

Table VIII and Fig. 7 show the goodness-of-fit indices and the path diagram, respectively. The path diagram and path coefficients show that *Cal_bas* has the greatest influence from *Alg*, and *Vec* has the greatest influence from *Fun*. From these results, it was found that the influence on these items showed a different trend from the integrated data.

TABLE VIII: GOODNESS OF FIT INDICES OF THE IMPROVED MODEL FOR 2015 DATA

df	*p*-value	GFI	AGFI	RMSEA	SRMR
1	0.513	0.999	0.978	0.000	0.0104

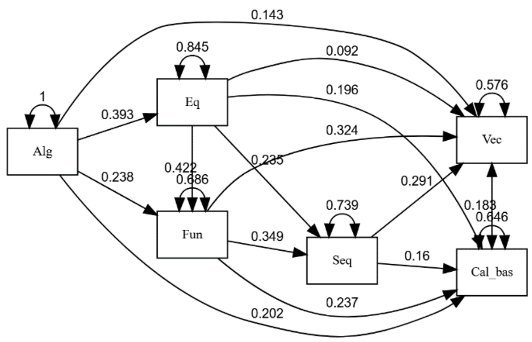

Fig. 7. Path diagram of the improved model for 2015 data.

VI. CONCLUSION

We analyzed the score data of the mathematics achievement exam equally scored out by learning item in mathematics, and created the path diagrams using the "sem" and the "pathDiagram" functions of the statistical analysis software R.

First, by considering the correlation values and the order of learning items, two models with the goodness-of-fit indices for the integrated data were obtained. As a result of analysis using these two models, it was found that the leaning items that had a great influence on later learning items were found. Note that similar results were obtained from the two models. Specifically, *Alg* and *Fun* have the most influence on *Vec* and *Cal_bas*, respectively. These influences are considered to come from the relationship between the following learning contents:

1) Algebraic calculation methods of expressions including variables and powers and radical sign, which are learned in *Alg*, are strongly linked to vector operations learned in *Vec*.
2) The content about functions learned in *Fun* is the fundamental for differentiation and integration of functions learned in *Cal_bas*.

Next, we applied these two models, which were created with data from multiple years, to single-year data, we confirmed that they fit well. We also analyzed the 2014 data, which was particularly well-fitted by model_1. Although the same model was used, different analysis results were obtained from the integrated data, indicating that this model can also analyze the characteristics of specific years.

Finally, although the 2015 data did not show the high goodness-of-fit index values for the two models, we were able to create another good fit model for this year. Analysis based on these models is planned to be used to improve the effect of mathematics education for technical college students, and we would like to continue further analysis.

In the future, we plan to create the following models to increase applicability: create a model that incorporates not only the learning items but also the total score as an observed variable; analyze students by dividing them into several groups according to criteria such as total score and variance; incorporate latent variables found by some method from observed variables.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Akio Ishida designed and conducted the research and drafted the paper; Naoki Yamamoto contributed by providing advice and suggestions for improving the research and checked the manuscript; Jun Murakami contributed as the supervisor and checked the manuscript; all authors had approved the final version.

FUNDING

This work was supported by the JSPS KAKENHI (Grants-in Aid for Scientific Research) under Grant No. JP21K13610.

REFERENCES

[1] M. Chen, S. Mao, and Y. Liu. "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, January 2014.

[2] A. Ishida, K. Kawakami, D. Furushima, N. Yamamoto, and J. Murakami, "Analysis of relationships between amount of physical activity of patients in rehabilitation and their ADL scores using multidimensional PCA," *Advances in Intelligent Systems and Computing*, vol. 690, pp. 147-158, November 2017.

[3] N. Oishi, A. Ishida, N. Yamamoto, and J. Murakami, "A causal analysis by structural equation modeling of sleep monitoring sensor data," *International Journal of Electronics and Electrical Engineering*, vol. 8, no. 3, pp. 58-62, September 2020.

[4] A. Ishida, N. Yamamoto, J. Murakami, and N. Oishi, "Solving 3-D puzzles using tensor decomposition and application to education of multidimensional data analysis," *International Journal of Machine Learning and Computing*, vol. 8, no. 5, pp. 447-453, October 2018.

[5] N. Yamamoto, A. Ishida, N. Oishi, and J. Murakami, "Development of teaching tool for supporting understanding of tensor decomposition using macmahon's coloured cubes," *International Journal of Information and Education Technology*, vol. 10, no. 1, pp. 14–19, January 2020.

[6] N. Yamamoto, A. Ishida, K. Ogitsuka, N. Oishi, and J. Murakami, "Development of online learning material for data science programming using 3D puzzle," *International Journal of Information and Education Technology*, vol. 11, no. 4, pp.154-163, April 2021.

[7] S. Abe, A. Ishida, J. Murakami, and N. Yamamoto, "Development of online learning materials for tensor data processing exercises," *International Journal of Information and Education Technology*, vol. 12, no. 3, pp.194-202, March 2022.

[8] *Guidelines of Management of Teaching and Learning, The Central Council for Education, Ministry of Education, Culture, Sports, Science and Technology*, Japan, 2020 (in Japanese).

[9] *Model Core Curriculum -Guidelines-(Excluding Economics and Business)*, National Institute of Technology, 2017.

[10] R. Core Team, "R: A language and environment for statistical computing," *R Foundation for Statistical Computing*, Vienna, Austria.

[11] L. Hu and P. M. Bentler, "Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification," *Psychological methods*, vol. 3, no. 4, pp. 424-453, 1998.

[12] R. A. Kusurkar, T. J. T. Cate, C. M. P. Vos, P. Westers, and G. Croiset, "How motivation affects academic performance: A structural equation modelling analysis," *Advances in health sciences education*, vol. 18, no. 1, pp. 57-69, 2013.

[13] A. Nazim and S. Ahmad, "A comparison between ordinary least square (OLS) and structural equation modeling (SEM) methods in estimating the influencial factors of 8th grades student's mathematics achievement in Malaysia," *International Journal of Scientific & Engineering Research*, vol. 4, no. 7, pp. 717-722, 2013.
[14] SEM. [Online]. Available: https://cran.r-project.org/web/packages/sem/sem.pdf
[15] DIAGRAMMER. [Online]. Available: https://cran.r-project.org/web/packages/DiagrammeR/index.html

Akio Ishida received the M.S. and Ph.D. in science from Kumamoto University, Japan, in 2010 and 2014. He is currently an assistant professor of the Faculty of Liberal Arts, Kumamoto College, National Institute of Technology, Japan. His research interests are in the area of multi-dimensional data analysis and numerical calculation.

Naoki Yamamoto received the Ph.D. in engineering from Kyushu Institute of Technology, Japan, in 2001. He is currently a professor of the Faculty of Electronics and Information Systems Engineering, Kumamoto College, National Institute of Technology, Japan. His research interests are in the area of multidimensional data analysis and numerical calculation. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and Kyushu Society for Engineering Education.

Jun Murakami received the Ph.D. in engineering from Toyohashi University of Technology, Japan, in 2000. He is currently a professor of the Faculty of Electronics and Information Systems Engineering, the Director of Library, Kumamoto College, National Institute of Technology, Japan. He is also a Board Member and Planning Chairman of the Research for Innovation and Synthesis of Technology in Kumamoto. His research area of interests includes statistical analysis, numerical calculation, and digital signal processing. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Information Processing Society of Japan (IPSJ), and the Human Interface Society Japan (HISJ).