IIAI Open Conference Publication Series
IIAI Letters on Business and Decision Science
Vol. 001, LBDS040
DOI: https://doi.org/10.52731/lbds.001.040

Impact of Missing Data on the Processing of Educational Questionnaires and the Reported Results

Hiroyuki Maruyama [*], Takaaki Hosoda [*],
Tokuro Matsuo [*]

Abstract

Recent developments in Information and Communication Technology have made it possible to utilize various types of data to improve methods and outcomes in the field of education. However, data verification requires proper handling of that data. Therefore, in this study, we examined missing data and the handling of information from an educational questionnaire of student perceptions about blended learning at a graduate school in Tokyo, Japan. The results suggest that the missing data in the studied questionnaire occurred in a specific cluster. However, in this case, the missing data were not significant enough to alter conclusions based on data analysis of the questionnaire results.

Keywords: Questionnaire, Education, Missing data, Statistics

1 Introduction

In recent years, there has been a series of changes in the education field in Japan. For example, the digital transformation (DX) of education data continues to lead to innovation and other changes in educational methods [1]. In addition, Institutional Research (IR) has been introduced to support decision-making in university management [2,3]. Various system changes and capital investments related to DX and IR have already been implemented.

Through these activities, the utilization of data for education is progressing. Simultaneously, using DX and IR also raises questions about how to handle the data appropriately. Processing survey data usually requires various forms of data cleaning, such as processing typos and outliers [4]. Although there are abstract guidelines about what kind of processing should be performed, they lack sufficient detail about how, specifically, to perform the processing.

Therefore, the purpose of this study is to examine data cleaning in educational data, determine an appropriate processing method for missing values, and clarify policies regarding such processes.

2 Previous Research

In data processing, three generation mechanisms are assumed concerning missing data [5]. The first is termed “missing completely at random” (MCAR), referring to missing data that does not depend on any particular variable and occurs completely randomly. The second term, “missing at random” (MAR), describes missing data occurrences related to another variable or variables. Third, “missing not at random” (NMAR) describes occurrences of missing data that depend entirely on an unobserved variable. Based on these established circumstances, various verifications

[*] Advanced Institute of Industrial Technology, Tokyo, Japan

have been developed for use with missing data. For example, Nissen, Donatello, and Van Dusen, who compared complete case analysis with multiple imputations, recommend multiple imputation analysis, having found greater bias in full case analysis [6].

Bias is further related to the case for restoring missing data. There are two merits of restoring missing data: suppression of bias and increased sample size. Concerning bias, survey respondents who leave some items unanswered may have unique preferences. If they are unable to contribute an answer for inclusion in the analysis, the analysis could lead to a false conclusion. Regarding increased samples, omitting respondents who do not respond excludes them from the analysis target. However, by making corrections, the respondents on the left can also be included in the analysis target.

We conducted this study in view of making an academic contribution to establishing the most appropriate method for handling educational data analysis.

These studies examine the appropriate preparation of data in education research. As such, it helps decision makers make appropriate behavioral decisions from the analysis of data when making educational decisions.

3 Analytical Method

3.1 Example

To The data examined in this study are from a student questionnaire conducted at the Advanced Institute of Industrial Technology, Tokyo. The survey was designed to assess students' awareness of blended learning. Blended learning represents a lesson format that combines face-to-face lectures with classroom lectures, online lectures, and video lectures. Compared to typical individual lesson styles, combinations of multiple lesson styles allow for more effective lesson management. The student questionnaire contents covered the following topics:

- Expectations for blended learning
- Expectations for recorded lessons
- Expectations for face-to-face lessons
- Purpose of studying at university
- Students’ feelings after taking the class

Students’ response choices were indicated on a five-step scale or the response “no experience.” The target participants were students attending the Advanced Institute of Industrial Technology. It was also conducted by an internet survey. The sample was 154 (recovery rate was 61.4%). When the questionnaire was conducted, the handling of personal information was explained and the consent of the subjects was obtained.

3.2 Method

The following processing was carried out for the conducted questionnaire. First, each question was divided into two groups: those who answered "no answer" or "did not experience" and those who answered other than that. Next, the test of the difference in the mean value between the two groups was performed for all the questions except those included in the two groups (above). Finally, the missing values were complemented by the regression imputation method: First, the ordinal logistic regression model defined in Equation (1) is estimated using the variables that had a significant difference in the test of the difference in the mean values among the data without missing data. Next, the values of the missing variables are complemented using the estimation results. Test the difference between the mean values for the original data and the complemented data. This makes it possible to verify the effect of performing complementation.

$$\log\left(\frac{\pi_k}{1-\pi_k}\right) = b_{0k} + b_1x_1 + b_2x_2 + \cdots + b_px_p. \tag{1}$$

Ordinal logistic regression analysis is a regression analysis for ordinal variables. Here, the order is estimated using p explanatory variables for the ordinal variables in n stages. In addition, π_k represents the probability of ordering up to the kth, b_i represents the coefficient, and x_i represents the explanatory variable.

4 Results

4.1 Result of Analysis

Table 1 summarizes the number of participants who answered normally and those who answered “no” or “no experience” in the questionnaire. However, the items described here are only items for which some people either did not answer or responded that they “did not experience it.” There were 10 questions with no answers or answers indicating they did not experience the given situation.

In the responses to these 10 questions, the number of questions with a significant difference was as follows:

- 5 questions: Gaining a deeper understanding through recorded lessons
- 5 questions: Watching recorded lessons becomes tiresome
- 5 questions: Watching recorded lessons will increase your motivation to study
- 4 questions: Recorded lessons can be watched repeatedly, which is useful for better learning.
- 4 questions: Good ccommunication with teachers
- 4 questions: Good ccommunication with other students
- 4 questions: It is difficult for me to make my opinion known in online face-to-face lessons
- 4 questions: You find it difficult to speak in online face-to-face lessons
- 4 questions: You can more easily express your opinions in online face-to-face lessons than in-class face-to-face lessons.
- 3 questions: You don't want to show your face in an online face-to-face class

Table 2 shows the results of a test of the difference in mean between the group that normally answered these 10 questions and the group that answered “no” or “no experience.” Of the 10 questions, some of the questions were unanswered by the same person. The following three questions are answered by same unanswered people:

- Can you gain a deeper understanding of recorded lessons?
- Do you get tired of watching recorded lessons on the way?
- Do you motivate yourself to study by watching recorded lessons?

Table 1: Number of Samples with Missing Data

Survey Item	**Number of samples**	**Normal answer**	**No answer or "I did not experience"**	**Answer pattern**
Gain a deeper understanding with recorded lessons	154	148	6	1
Watching recorded lessons gets tired on the way	154	148	6	1
Watching recorded lessons will increase your motivation to study	154	149	5	2
Recorded lessons can be watched repeatedly, which is useful for better learning	154	148	6	1
Communicate well with teachers	154	148	6	3
Communicate well with other students	154	148	6	3
It is difficult for me to make my opinion known in online face-to-face lessons	154	147	7	4
I Find It Difficult to speak in online face-to-face lessons	154	144	10	5
It is easier to express opinions in online face-to-face lessons than face-to-face lessons in the classroom	154	149	5	6
I don't want to show my face in an online face-to-face class	154	148	6	3

In addition, the same respondent did not answer the following questions:

- Can you communicate well with teachers?
- Can you communicate well with other students?
- Do you find it difficult for your opinion to be reflected in online face-to-face lessons?

Therefore, as a test of the average value, since the results are the same for the questions of the same non-answer pattern, they are summarized in 6 patterns. As a result of the test, there was a significant difference in a total of 11 questions.
Table 3 shows the test results of the answers that complemented the answer that there was no answer or that they did not experience it. For complementation, estimation is performed using questions that have a significant difference for each question. There were no questions with significant differences in the mean test.

Table 2: Complementary Results of Missing Data

Survey Items			**What are your expectations for the following in all blended classes?**			**What are your expectations for the following in recorded lessons (videos) in blended lessons?**	
			Link of knowledge of recorded lessons and face-to-face lessons	***Opportunity to communicate with teachers by email outside the classroom***	***Presenting a large amount of preparation/review tasks***	***Humorous explanations in recorded lessons***	***The teacher's facial expression can be seen in the recorded lesson video***
1	Gain a deeper understanding with recorded lessons Watching recorded lessons gets tired on the way	Normal answer	4.176			3.547	
	Watching recorded lessons will increase your motivation to study	"No answer" or "I did not experience"	4.667			4.333	
2	Recorded lessons can be watched repeatedly, which is useful for better learning	Normal answer	4.174				
		"No answer" or "I did not experience"	4.8				
3	I communicate well with teachers I communicate well with other students	Normal answer	4.176		2.723		
	It is difficult for my opinion to be reflected in online face-to-face lessons	"No answer" or "I did not experience"	4.667		3.667		
4	It is difficult to speak in online face-to-face lessons	Normal answer					
		"No answer" or "I did not experience"					
5	It is easier to give opinions in online face-to-face lessons than face-to-face lessons in the classroom	Normal answer		3.542		3.528	3.264
		"No answer" or "I did not experience"		4.1		4.3	4.3
6	I don't want to show my face in an online face-to-face class	Normal answer			2.738		
		"No answer" or "I did not experience"			3.4		

Table 2: Complementary Results of Missing Data (Continue)

Survey Items			**What is the purpose of studying at this university?**					
			Career formation through degree acquisition	***Increase friends and acquaintances***	***Escapism***	***Orders from the company or boss***	***For entrepreneurship***	***To go on to a doctoral course***
1	Gain a deeper understanding with recorded lessons	Normal answer				1.189	2.399	2.818
	Watching recorded lessons gets tired on the way							
	Watching recorded lessons will increase your motivation to study	"No answer" or "I did not experience"				1	1.333	1.833
2	Recorded lessons can be watched repeatedly, which is useful for better learning	Normal answer			1.792	1.188	2.389	
		"No answer" or "I did not experience"			1.2	1	1.4	
3	I communicate well with teachers	Normal answer				1.189	2.392	
	I communicate well with other students							
	It is difficult for my opinion to be reflected in online face-to-face lessons	"No answer" or "I did not experience"				1	1.5	
4	It is difficult to speak in online face-to-face lessons	Normal answer	4.027	3.51		1.19	2.395	
		"No answer" or "I did not experience"	4.714	4.143		1	1.571	
5	It is easier to give opinions in online face-to-face lessons than face-to-face lessons in the classroom	Normal answer		3.507				
		"No answer" or "I did not experience"		4				
6	I don't want to show my face in an online face-to-face class	Normal answer				1.188	2.389	
		"No answer" or "I did not experience"				1	1.4	

Table 3: Complementary Results of Missing Data

Pattern	**Question**	**Original data**		**Complemented data**		***p* value**
		Average	***SD***	***Average***	***SD***	
1	Gain a deeper understanding with recorded lessons	3.838	1.167	3.844	1.144	0.962
	Watching recorded lessons gets tired on the way	2.905	1.280	2.903	1.256	0.985
	Watching recorded lessons will increase your motivation to study	3.709	1.051	3.701	1.036	0.946
2	Recorded lessons can be watched repeatedly, which is useful for better learning	4.423	0.894	4.442	0.886	0.855
3	I communicate well with teachers	2.791	1.191	2.786	1.171	0.972
	I communicate well with other students	2.392	1.260	2.377	1.237	0.915
	It is difficult to express my opinion in online face-to-face lessons	2.581	1.149	2.584	1.130	0.980
4	It is difficult to speak in online face-to-face lessons	2.973	1.260	2.974	1.231	0.993
5	It is easier to share opinions in online face-to-face lessons than face-to-face lessons in the classroom	3.118	1.372	3.136	1.334	0.907
6	I don't want to show my face in an online face-to-face class	3.570	1.291	3.584	1.272	0.925

4.2 Discussion

There were six patterns for the unanswered questions in Table II. In the first group, the value of "humorous explanation in recorded lessons" was high, and the value of "for going on to doctoral course" was low. This suggests the possibility of no answer in clusters that are more interested in lesson content than career development. In the second group, the values of "link of knowledge between recorded lessons and face-to-face lessons" and "escape from reality" were low. This suggests the possibility of no answer in a cluster with a clear learning purpose. In the third and sixth groups, there was a high expectation that "a large amount of preparation/review tasks would be presented." This suggests the possibility of no answer in the cluster that seeks the amount of learning. In the fourth group, the values of "career development through degree acquisition" and "increasing friends and acquaintances" were high. This suggests the possibility of no response in clusters aiming to expand their careers and connections. In the fifth group, "Opportunities for communication with teachers by email outside the classroom," "Humorous explanations in recorded lessons," "The facial expressions of teachers can be seen in the video of recorded lessons," and "Friends/The value of "Increase acquaintances" were high., suggesting the possibility that the clusters of no responses were in items that emphasize communication and connections.

However, as a result of the complement shown in Table III, no group had a significant difference from the original data. This suggests that, as shown in Table I, there were few unanswered responses; hence, there was no significant difference in the overall trend. However, the supplemented answers were almost the same. Therefore, when the number of unanswered samples increases, the overall tendency may change.

5 Conclusions and Future Research

In this study, we verified missing data in a questionnaire in the field of education. The results show that the missing data may belong to a particular cluster. However, as a result of the complementation, there was no significant change from the original data. However, in this study, the small number of missing samples may have an effect. Therefore, it is necessary for future studies to verify how much missing data changes the conclusions obtained from the data. In addition, comparison of missing data complementation methods and application to broader education data, not limited to questionnaires, are issues we recommend for further study.

References

[1] R. R. Puentedura, "SAMR: Moving from enhancement to transformation," http://www.hippasus.com/rrpweblog/archives/2013/04/16/SAMRGettingToTransformation.pdf, (2013)

[2] J. L. Saupe, The Functions of Institutional Research, Association for Institutional Research, 1990.

[3] J. F. Volkwein, "The four faces of institutional research," New Dir. Insit. Rese, Issue 104, pp. 9-19, 1999.

[4] E. Rahm, and H. H. Do, "Data cleaning: Problems and current approaches," IEEE Data Engin. Bull., Vol. 23, pp.3-13, 2000

[5] P. D. Allison, Missing Data, SAGE Publication, 2001

[6] J. Nissen, R. Donatello, and B. Van Dusen, "Missing data and bias in physics education research: A case for using multiple imputation," Phys. l Rev. Phys. Educ. Res., Vol. 15, Issue 2, 2019.