Choosing appropriate independent variable in educational experimental research: some errors debunked

R L Panjaitan¹,*
¹Pendidikan Guru Sekolah Dasar, Kampus Sumedang, Universitas Pendidikan Indonesia, Jalan Mayor Abdurrahman 211, Sumedang 45322, Indonesia

*Corresponding author: regina@upi.edu

Abstract. It is found that a number of quantitative research reports of some beginning researchers, especially undergraduate students, tend to ‘merely’ quantitative with not really proper understanding of variables involved in the research. This paper focuses on some mistakes related to independent variable determination in experimental research in education. With literature research methodology, data were gathered from an undergraduate student’s thesis as a single non-human subject. This data analysis resulted some findings, such as misinterpreted variables that should have represented the research question, and unsuitable calculation of determination coefficient due to incorrect independent variable determination. When a researcher misinterprets data as data that could behave as the independent variable but actually it could not, all of the following data processes become pointless. These problems might lead to inaccurate research conclusion. In this paper, the problems were analysed and discussed. To avert the incorrectness in processing data, it is suggested that undergraduate students as beginning researchers have adequate statistics mastery. This study might function as a resource to researchers in education to be aware to and not to redo similar errors.

1. Introduction
Quantitative research in educational field becomes more common these days [1]. However, using quantitative methods in education might bring some consequences due to statistics involvement. Researchers should be mindful of all included variables and know how to use statistics in data processing correctly and effectively.

Collected data should be confirmed to represent the variables existing in the research questions or in the research title. In quantitative research, researchers would like to know whether one of the variables of a predetermined value might estimate of the other [2]. Nonetheless, having a proper understanding of variable is one challenge many students encounter [3]. Researchers are probable to gather data that did not necessarily depict the related variable. This kind of errors would lead to incorrect conclusion. This inaccurate conclusion could delude researchers who would like to do following studies.

A simple survey was conducted by the author with 37 participants who were primary school teacher candidate program undergraduate students. The participants were given a statement, “When working on my thesis, I tend to copy the data calculation from my senior’s thesis” and they should decide how often they did it. Interestingly, there are 2 students (5.4%) answering “always”, 11
students (29.7%) said “often”, “sometimes” was admitted by 24 students (64.9%), and no one answered “never”. This survey shows that these students really rely their theses on their seniors’ theses. Therefore, it is intelligible to be aware that errors made by senior students, as previous researchers, might be derived by beginners in educational research.

Moreover, it is said that some statistical misconceptions are generated by improper statements that are found in books, seen on the Internet, and presented orally in classrooms [4]. It means that it is needed to decrease the misconceptions by reassuring that all of research reports (including thesis) that are being written or would be published have no statistical-concept errors.

This paper discusses examples of some improper variable determinations and calculations on a work of a thesis. This study is expected to be a resource for researchers so that they might avoid doing the same errors.

2. Methods
Literature research methodology was used in undertaking this work. Literature research methodology is to study and examine varieties of reading resources for pointing out crucial elements of information [5]. Data were gathered by collecting information from a subject which is a thesis of an undergraduate student of elementary school teacher education study program. The data were investigated with content analysis, which is a systematic approach in examining texts to obtain discoveries [6]. Thus, the findings were critically elaborated and any errors found in the results were informed to prevent other researchers from doing similar faults.

3. Results and discussion
Most of the discussion would be about an undergraduate student’s thesis with topic “The Effect of Problem Based Learning on Students’ Mathematical Understanding Ability” [7]. This thesis is a research report of an experimental research conducted with 31 participants in the control group and 32 participants in the treated group. The treated group is a group that experienced Problem Based Learning in the class, and the untreated group experienced traditional learning.

3.1. Case 1: Determining improper independent variable
The thesis author gathered data of the treated group as can be seen in Table 1.

| Sample | Pretest Score | Posttest Score |
|--------|---------------|----------------|
| S1     | 18            | 62             |
| S2     | 24            | 64             |
| S3     | 26            | 74             |
| S4     | 15            | 58             |
| S5     | 23            | 55             |

From the data as in table 1, the thesis author calculated the correlation between pretest score and posttest score using Spearman’s rho, and got result as in Table 2.

| Spearman’s rho | Pretest score | Posttest score |
|----------------|---------------|----------------|
| Correlation Coefficient | 1.000 | .420* |
| Sig. (2-tailed) | .017 | .017 |
| N | 32 | 32 |

| Posttest score | Pretest score | Posttest score |
|----------------|---------------|----------------|
| Correlation Coefficient | .420* | 1.000 |
| Sig. (2-tailed) | .017 | .017 |
| N | 32 | 32 |
From table 2, the thesis author concluded that the correlation coefficient between Problem Based Learning approach with students’ mathematical understanding ability improvement is 0.420 (positive correlation). Actually, it is needed to be really cautious that two variables that have been correlated are “Pretest Score” and “Posttest Score”, and nothing from those variables can be designated as the variable “Problem Based Learning” as the original dependent variable in the research title. Mathematically, correlation coefficient can be calculated from any two sets of data. However, a researcher should be alert to the variables the data would represent. The correlation $r = 0.420$ clearly stands for correlation between pretest and posttest score, and it means the previous conclusion saying that it is the correlation coefficient between Problem Based Learning approach with students’ mathematical understanding ability improvement is unsuitable.

3.2. Case 2: Calculating inaccurate coefficient of determination

As the continuation of Case 1, from Table 2 the thesis author inferred that the determination coefficient can be calculated with this equation [7].

\[
\text{coefficient of determination} = r^2 \times 100\% \quad (1)
\]

The correlation $r = 0.420$ in Case 1 is substituted to Equation (1), and the thesis author got a result that the coefficient of determination $= (0.420^2) \times 100\% = 17.64\%$. The thesis author concluded that Problem Based Learning approach (independent variable) has 17.64% contribution on students’ mathematical understanding ability improvement (dependent variable).

It is known that coefficient of determination is a measurement of adequacy of the regression model referring to the variability in the data [8] which value equals to the square of Pearson’s $r$ ($r^2$) [9]. Pearson correlation for variable $X$ and $Y$ is expressed as [9]

\[
r = \frac{\Sigma (X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma (X-\bar{X})^2 \Sigma (Y-\bar{Y})^2}} \quad (2)
\]

Meanwhile, the Spearman’s rho correlation coefficient formula that has been calculated by the thesis author is as follows [10].

\[
\text{Spearman} \ r = 1 - \frac{6 \Sigma (d_i)^2}{n(n^2-1)} \quad (3)
\]

where $d_i$ = the difference between ranks and $n$ = the number of pairs of $(X, Y)$ scores or the number of $d_i$ differences.

Spearman $r$ is one of correlation calculations for non-normal distribution data [10] where Pearson’s $r$ cannot be applied. Furthermore, it can be seen that equation (3) does not equal to equation (2).

Since equation (3) has different values from equation (2), and $r^2$ in equation (1) has the same result as the square of $r$ in equation (2), it is clear that the result of Spearman $r$ calculation cannot be perceived as $r$ in the calculation of coefficient of determination. Not all correlation coefficients in statistics can be simply considered as $r$ that can be used in equation (1).

Moreover, referring to Case 1, the dependent variable is already inappropriately chosen. The data that have been correlated are pretest score and posttest score on students’ mathematical understanding ability improvement, and reasonably pretest score does not represent Problem Based Learning Approach. This resulted that the following computation using pretest score as dependent variable is inaccurate, including the coefficient of determination and the conclusion that follows.

3.3. Data processing for calculating the effect of an approach in educational setting

Research entitled “The Effect of Problem Based Learning on Students’ Mathematical Understanding Ability” is an experimental research that can be conducted with a randomized pretest-posttest control group design that is categorized as one of true experimental designs. Two groups of subject (as treated and untreated groups) are used, with both groups being measured twice for pretest and posttest [11]. Random assignment is used in forming the groups. The mean of the gain of each group would be statistically compared (for some cases it could be calculated with independent-samples $t$-test) and
eventually it can be inferred whether the mean of the gain for each group is significantly different or not. If the gain mean is significantly different, it can be inferred whether the treated or untreated group that has a better effect of the learning approach.

Besides, if the calculation of coefficient of determination is needed, for some cases it can be computed by calculating correlation coefficient involving dummy variables. The researcher may use dummy variable for indexing untreated and treated group. Then, when it is considered that there is no difference in pretest-score means for both groups, posttest-score of the treated and untreated groups may act as the dependent variable (Y) and the dummy variables (0 for untreated group and 1 for treated group, for instance) act as the independent variable (X). Thus, the correlation coefficient can be examined. This test of correlation coefficient and the independent-samples t-test as mentioned previously are mathematically equivalent [4]. It is known that Pearson’s r can also be used to assess the strength and direction of relationship with dummy variables [9]. From this Pearson’s r value between the dummy variables as independent variable and posttest score (for both groups) as the dependent variable, the coefficient of determination as the square of Pearson’s r may be calculated.

4. Conclusion

From Case 1 in Discussion section it can be clearly noticed that using pretest data-set for independent variable does not represent the Problem Based Learning approach as it is needed in the thesis title. This inappropriate data-set determination as independent variable leads to inaccurate calculation of coefficient of determination and evidently results erroneous conclusion. Mathematically, almost all data can be paired or computed, but not all data can be simply interpreted as the independent variable which suits with the research topic. With this paper, it is expected that researchers could avoid doing similar errors.

For preventing inaccuracy in processing data, adequate statistics enrichment such as how to use it, what kind of data that is logical to be calculated with it, is needed by beginners in educational research, including students majoring education. Unfortunately, students at the undergraduate level tend to view research methods courses negatively [12]. A thorough and interesting statistic class involving examples which are correct data processing and incorrect ones might be undertaken.

In addition, educational researchers should always check their reference on data calculation and use common sense in research. Published work does not necessarily mean suitable-work-to-copy. To elude errors on data interpretation, researchers should be cautious on imitating others’ calculation.

References

[1] Göktaş Y, Hasançebi F, Varışoğlu B, Akçay A, Bayrak N, Baran M and Sözbilir M 2012 Educational Sci.: Theory & Practice 12 455
[2] Coşkuntuncel O 2013 Educational Sci.: Theory & Practice 13 2151
[3] Lucariello J, Tine M T and Ganley C M 2014 J. Mathematic Behavior 33 30
[4] Huck S W 2009 Statistical Misconceptions (New York: Psychology Press)
[5] Lin G 2009 Int. Education Studies 2 179
[6] White M D and Marsh E E 2006 Library Trends 55 22
[7] Respati R 2016 Pengaruh pendekatan Problem Based Learning terhadap Kemampuan Pemahaman Matematis dan Komunikasi Matematis Siswa pada Materi Skala dan Perbandingan. Undergraduate [thesis]. Sumedang: Universitas Pendidikan Indonesia
[8] Miah A Q 2016 Applied Statistics for Social and Management Sciences (Singapore: Springer Nature)
[9] Healey J F 2012 Statistics: A Tool for Social Research (Wadsworth: Cengage Learning)
[10] Warner R M 2013 Applied Statistics (Washington DC: Sage Publications, Inc.)
[11] Fraenkel J R, Wallen N E and Hyun H H 2012 How to Design and Evaluate Research in Education, 8th edition (New York: McGraw-Hill)
[12] Papanastasiou E C 2005 Statistics Education Research J. 4 16