Applying Knowledge of Probability Statistics in Science Research and Practice

Author Details: Le Hang Le
University of Economics - Technology for Industries, Vietnam
Correspondence: **Le Hang Le**, 456 Minh Khai, Hai Ba Trung, Ha Noi

Abstract:
The overview article introduces the application of the subject of statistical probability in scientific research and in practice business activities of Vietnamese enterprises. Statistical probability is an extremely important science in scientific research. From statistical probability parameters that help researchers predict trends and possibilities and give the degree of impact of factors on a research topic. In business practice, statistical probability is also an effective tool in decision making and price control as well as evaluating the business performance of the business.

Keywords: *Probability statistics, scientific research, Vietnamese enterprises*

1. Introduction

Educational science research (academic science research) is an urgent need to build a theoretical basis to lead the way for improvements and reforms in education, and the results of serious research on educational science are an important basis for educational decisions for a country, avoiding the possible consequences of a "trial-and-error" approach. Knowledge of educational science is not only needed for leaders and policy-makers, but also for all those working in the education sector, because they are the builders of the educational system. In most advanced countries in the world, scientific education research is interested in investment and development, has made great achievements and made effective contributions to the development of national education. .

In the general context of the integration period, our Vietnam, including pedagogical universities, is increasingly paying more attention to the field of pedagogical research, besides the fields of basic scientific research, because pedagogical research is a the first and foremost important and practical field for the quality of the school's training activities. In parallel with the promotion of research on educational planning, fostering to improve research capacity for faculty members is also focused, including the content of applying mathematical statistics to quantitative studies in Business planning. In fact, this is a weakness and a real fostering need of a part of lecturers at present pedagogical universities. This weakness is one of the important reasons that make our scientific research results difficult to be accepted for publication in specialized journals abroad.

In the late 20th century, when it comes to the quality of educational scientific researches, many researchers inside and outside the education sector also have a lack of confidence in the accuracy or convincing of the studies. This compares with studies in the natural sciences such as Physics or Chemistry ("hard science"). This led Larry Hedges to conduct research in 13 areas of Psychology and Education ("soft science") and compared with research in Physics. As a result, Hedges found a similar situation occurring both in the natural sciences and in the social sciences, that approximately 50% of the studies showed different results in both Psychology. , Education and Physics. That is, studies in Physics also give conflicting results similar to those in education. This is because researchers often dismiss studies that tend to produce extreme results when combined. Even more so in natural sciences than in social sciences (40% versus 10%). These findings were presented by Hedges in a report titled “How Hard Is Hard Science? How Soft Is Soft Science? ”. In which he made the general remark that research in the "soft sciences" can be compared to research in "hard sciences" in terms of accuracy and reliability. Educational science researchers as well as other researchers may find general trends in the findings from long-term studies. There should be as many studies as possible on the same topic and a synthesis and analysis of those studies. The combined results of these findings will be considered to be the best possible assessment of what is known on the subject [lead to 1]. Hedges' suggestion affirmed the role of applying mathematical statistics to pedagogical research, because in order to do that, we must use quantitative methods and techniques.

2. Applying statistical probability in scientific research

2.1. Characteristics of statistical probability in scientific research

One of the basic characteristics of the science of the present world is that the sciences increasingly use methods of mathematics. Mathematics has penetrated the fields of: Biology, Medicine, Linguistics, Psychology ... Recently, mathematics has entered the educational science with statistical probability theories, information theory, math logic ... and has yielded clear results. The tendency of mathematics has opened up new ways, deepening into the nature and law of research phenomena as C. Marx said, a science can only really develop if it can use math. learn.

Educational science research has the common characteristics of scientific research as a creative activity, proposing new things and towards discovering the truth, so there are strict requirements for the researcher. . It is the requirement of objectivity and accuracy in research, expressed in fidelity to objective reality while discovering a new one without further modifying it according to the subjective wishes of the researcher or of the researcher. other one. The research object of educational science (education science) is phenomena, processes are very complex, always fluctuate due to the influence of many objective and subjective factors, so there will be a series of factors. need to be controlled in the process of researching. Objective and accurate requirements, first of all, require the selection of research methods, measures, tools and techniques to be less affected by subjectivity. The more reliable the researcher of the mediator is, the more reliable it is. Because scientific research results as a basis for making decisions about education and educational development. Without credibility, it will greatly affect the feasibility of the given policies, which will affect the entire social life.

The processing of collected information is an extremely important stage in the whole process of organizing, implementing and implementing a scientific research topic. This is because the collected facts, documents, and information are not the resolution of research tasks by themselves. They need to be properly summarized, analyzed, interpreted and generalized. This is done by processing the obtained documents both qualitatively and quantitatively. The purpose of qualitative analysis is to establish the different qualities and properties of the phenomena under study. When analyzing qualitatively, we can use known indicators and determine whether or not they are present in trials. In the education plan qualitative issues are mainly. However, the quantifiable problems will help the qualitative become more specific and accurate because the typicality of the quality to be studied can be clarified. Of course, there are aspects in education which can only be conventionally quantified and many aspects that cannot be quantified. But if you want to draw a larger practical application, you need to try to quantify.

Thus there are many ways to process the information gathered, and mathematical statistics processing is an important method.

The volatile educational phenomena make it practically impossible to control all the errors in a study, affecting the accuracy of the study. For example, an educational researcher cannot do two experiments under completely identical conditions because in reality it is not possible to have 2 students alike in all aspects, nor can there be 2 classes with perfect conditions. all the same. Experiments in pedagogical research are random experiments, but random does not mean messy, not according to the rules. It is not possible to accurately predict the results of individual experiments, but if one moves from individual experiments to a series of random experiments under certain conditions, even the individual results do not a rule, but the mean results of many randomized experiments are consistent. That is why researchers assign their results a measure of probability. When researchers report their results as significant at the 0.05 level, it means that there is only 5 out of 100 chance that their results would have uncontrolled errors in the study. When researchers reported that their results were significant at 0.01 it was only 1 in 100 that their results were uncontrollable. By combining the results of multiple studies, it is possible to draw much more solid conclusions than when there was a single study.

Probability theory is the science of the laws of random phenomena, is an effective tool for scientific education.

Mathematical statistics is a part of probability theory, whose research object is to collect and summarize observational data, experiment, analyze to draw reliable conclusions from those data. Education often requires processing a huge amount of data such as: number of students, number of teachers, learning results of students ... Mathematical statistics provide a way to summarize data to monitor the situation, help investigate and evaluate the quality of education, compare the effectiveness of two educational methods, analyze the relationship between educational phenomena, analyze the effects of factors on a educational phenomenon ... In this article, we will mention a few specific issues in order to clarify the significance and effects of mathematical statistics in the study of educational science.

Quantification in pedagogical research depends firstly on the most fundamental issue of how to correctly and accurately (by number) the characteristics of the research phenomena. This is a complicated problem because, as mentioned above, not all educational phenomena can be accurately quantified at present.

2.2. Some basic applications of mathematical statistics to educational science research

*The scale

In order to be able to use mathematical tools to process the obtained data, first of all, it is necessary to properly solve the fundamental problem of measuring or quantifying the characteristics of the research objects. closely related to the concept of "measurement". It is the comparison of a certain quantity with a known reference, and the result is to give the numbers to evaluate [3]. Thus, "Measurement" is to assign the objects and its properties to be measured according to defined principles. These rules specify the conformity between some characteristics of the numbers with some of the features to be measured of the object. Depending on the degree to which compliance is determined, the following measurements may be made (and their respective scales):

-The measurement "Identity" (Nominal - or type) is to separate a certain sign of the feature to be studied and marked every time that sign is encountered (in observation or in experiment). Aggregating (counting) the numbers recorded will have a characteristic manifestation of the subject or research phenomenon (for example, it is possible to evaluate students' ability to write correct words by recording the number of errors they make. must be in a spelling). The test identifier is the "Scale of Identifier".

- The "Ordinal" measurement is to rank the phenomena, the objects of the studied feature into a series in descending or ascending order, and then assigning each object a number, which that number specifies the object's position in the sequence. This number is called the rating of the object (for example, the student's rating in a class - "first", "second", "three" ... based on learning performance in a subject). rating is "rating scale" (or scale).

- "Interval" measurement (Interval - exact measurement) is the comparison of research features with standard measurement units. Corresponding to the interval measurement there is a "interval scale".

- The measurement of "scaling" (ratio). Corresponding to it there is "Scale Scale".

The general development of science and information technology in recent years has helped metrology have strong developments, making measurements more and more accurate, and more and more sophisticated units of measurement.

Accurate measurement (approximate measurement) by comparison with units of measurement is limited in its application to pedagogical research because units of measurement are required, and measured characteristics do not change over the course of time. measurement time. These conditions are generally difficult to satisfy with the ever-changing phenomena and processes in education. Therefore, in pedagogical scientific research, two commonly used quantifiers are "Identifier" and "Qualifier". For example, when there is no direct measure of the quality of knowledge and skills, or the level of development of a student's ethical qualities, it is possible to mark the superficial manifestations of the student's behavior, errors, the performance results ... From there quantify those characteristics and discover the corresponding laws. Although we have not exactly determined, that feature A has a feature several times that characteristic in object B, but we can determine that that feature in A

is "Stronger", "More developed" in B. Therefore, if the research objects can be ranked, it will be possible to use the appropriate tools of mathematical statistics to find reliable conclusions. Depending on the defined scale, statistical parameters will be calculated (median, correlation coefficient, or arithmetic mean, variance, linear correlation coefficient ...).

For a 5-step or 10-step scale score that measures a student's level of knowledge, it is neither an interval measurement, nor a grade measure. In many cases, ranking signs are assigned. types of student assignments 0 to 10 or 1 to 5. This scoring is a combination of markup and ranking, often used in precise subjects or objective tests. If considered strictly, for the scores it is impossible to calculate the parameters: arithmetic mean, variance, linear correlation coefficient ... However, in practice, it is possible to approximate processing. The scores such as numbers are obtained according to the measurement range.

* Some common quantitative methods are now widely used in pedagogical research

The first content of mathematical statistics is the description of the observed results, the experimental results, ie trying to summarize a large number of data into a small number of features expressed in the form of a common concentration. Information contained in the data to help solve more fundamental problems: Scientific analysis of the data, thereby drawing generalized conclusions, drawing laws.

Usually, we want to learn the following:

The general trend or characteristics of the data collected;

- Differences in the figures obtained.

The statistical requirements are:

- Describe what is typical for a set;

- Describe the size or variability of the population;

- Describe the relationship between two variables in a set;

- Determine the probability of a completely random statistical quantity [4].

The first thing in summarizing the data is to look at the frequency distribution of the results obtained: Once the raw score has been obtained, the first thing is to sort the scores and record the number of occurrences corresponding to each score. numbers from high to low into one column. Then add all of them to get the frequencies. From that series of data summarized into a frequency distribution table. In order to make a more general assessment of the situation, those figures will be combined into classes or classifications (for example, classifying students with scores 8,9,10 in the "fair" and "excellent" categories. "; Hs with score 5,6,7 are classified as" average "; those students with score 4 or less are classified as" weak "and" poor ".

In the study of educational planning, people often apply the following groups of formulas for calculating statistics:

a / Group of good central positions to describe generally a set of scores. In this group there are centering measures. The purpose is to measure the mean or typical values of a population.

Average is a familiar parameter, characteristic of data concentration, used to reflect the mean of the distribution measured in scale or range.

Median is used for the variables measured by the rating level (rank).

The factor is used for variables measured at the rating level. The factor is the number at which the frequency is greatest. Therefore, in practice there can be a dual flavor or no weak flavor.

Thus, there are many mean scores, but people often think there is only the mean, so the mean is the most widely used. This is because it is clear, easy to calculate and is an algebraic function of all values of a variable.

With a sufficiently large number of observations, it reliably evaluates the parameter of the statistical population. But there are also cases where the average is not calculated. That is:

- When the number of observations (n) is too small;

- When the distribution is too asymmetric;

- When the distribution has an open layer at the end;

- When the distribution has multiple peaks.

The factor numbers are used in the distribution case with multiple vertices. Then the number of members has a clear meaning and can be easily determined. But it is not a function of all variable values, and is uncertain (varies widely from experiment to experiment). If the distribution is very symmetrical, then mean = mean = number of factors.

b / Group of measurements of the variable value of the data

Standard deviation (SD). SD is a value of the variability. It indicates the dispersion of the distribution of values: the larger the deviation, the more the dispersion, and vice versa, the smaller the deviation, the less the dispersion.

- In the case of two data sets with different mean values, the dispersion of the figures is compared by the coefficient of variation V, that is the ratio between the standard deviation and the mean. addition of the data sheet (usually calculated in%).

c / Group of measurements of relationships between variables participating in the study

Correlation coefficients (Correlation Coefficient) are calculated to find out the relationship between variables or factors. This is very important because it allows researchers to understand and understand the characteristics and nature of the phenomenon being studied. From there, find the best path to conduct educational impact.

Correlation coefficient is used to represent the correlation between two or more sets of values in two or more different distributions, or the correlation between factors involved in an experiment. There is positive and negative correlation (ranging from +1 to -1). There is also a straight correlation and a curved correlation.

The correlation coefficient can be calculated according to the following mathematical formulas:

* Pearson linear correlation coefficient (R) is also known as the product moment correlation coefficient. It is used when both variables are measured by the interval. This is the linear correlation coefficient between two variables (positive or negative. For example, find the correlation between the result of knowledge and attitude at the posttest point of the experimental group). Pearson's correlation coefficient cannot be used for curvature correlations.

To see if the correlation has statistical significance, it is necessary to calculate the T-student value

* Spearman hierarchical correlation coefficient (R) is used when both variables are measured in rank (rank).

The purpose of this correlation coefficient is to find correlation between two variables with different measures (for example, find correlation between the result of knowledge and attitude at the posttest score of the experimental group). The sample must be greater than or equal to 30.

Like tests using hierarchical levels, the Spearman's correlation coefficient is not concerned with the value of the score, but rather their hierarchical relationship in the score set. Therefore, when applying tests at this level of measurement, it is necessary to know how to chart the numbering.

In summary, the correlation coefficient is used to calculate in the following cases:

- Find the relationship between a physical characteristic with a psychological feature of a group of people (Pearson's correlation coefficient);

- Find a relationship between 2 certain psychological characteristics of a group of people (Pearson's correlation coefficient);

- Find a relationship between the ability of a group of students in one subject to compare with another (Pearson's correlation coefficient);

- Investigate the validity and reliability of the test (Pearson correlation coefficient);

- Find correlation between two related groups on a certain characteristic of Spearman's ecology);

- Find correlation between 2 variables with different measurements (Spearman correlation coefficient).

In the case of calculating the Spearman correlation coefficient, the number of individual samples must be greater than or at least equal to 30.

d / Group of measures for the difference between the variables

The statistical laws that specify the difference between the set of points is sometimes the result of the effects of random factors. Therefore, it is necessary to use statistical mathematical tests to confirm whether the observed change is significant or not. In other words, it is statistically significant or not. When making value assessments, there can be two types of mistakes: "bad" but rated as "good" and, conversely, "Good" is rated as "Bad". From a practical point of view, one should avoid making the first type of mistake and accept some false probability of a type 2 mistake because of less harm. According to the above spirit, when evaluating the effectiveness of the educational impact or the relationship between the two variables of attitudes and knowledge, we should derive from the assumption that: educational impact is ineffective (measured results of experimental class is not different from control class, posttest results are not different from pretest results in experimental class, there is no relationship between knowledge and attitude). Such a hypothesis is called a "zero hypothesis", and it has become a unified convention in mathematical statistics. If you can prove that the null hypothesis (Ho) is only true with the probability of 5%, you can safely reject the null hypothesis and accept the hypothesis H1 (effective, related) because then the false probability just 5%. When rejecting the Ho hypothesis, accepting hypothesis H1 with a 5% error probability, we say the found value is "significant in statistical probability". As mentioned in the first part, in scientific research, the accepted error is 5%. If we want to be more certain, we can accept the value to reject the hypothesis Ho with a 1% false probability (p is less than or equal to 0.01). The result will be called "Nonsignificant" if the probability of error is greater than 5% (P greater than 0.05). We can also reject the null hypothesis Ho with greater error, for example when P is less than or equal to 0.1. When accepting such hypothesis H1, the probability of error will be 10%, that is, in 100 confirmed cases, there are about 10 times of false.

Finding the differences between the variables is done through the mean or variance. The aim is to find a significant difference between groups categorized by a certain variable. In this heading there may be the following measurements:
* Compare the difference in the mean of a group (knowledge, attitudes: pretest and posttest. For example, evaluate the effect of population education on knowledge of first year students after a real year. test) to demonstrate the effectiveness of the experiment. To do this we need to:
- Determine the average value;
- Determination of standard error value;
- T-student dual wave group ("Double wave group" is a statistical term that shows the set of points of two experimental groups that have the same relationship, such as: husband-wife, brother-brother ... pretest-posttest matching of the same group).
* Compare the set of posttest scores of the experimental group and the control group / or two experimental groups with 2 different methods to find the method more effective.

In this case a non-double wave T-student test is often used with the following conditions:
- The two groups are not related to each other (completely independent of each other);

- The facts are equivalent (except for impact conditions);
- The best sample should be equal (N).

* Compare the results between the experimental groups and the control by Chi-square test.

The purpose is to compare the results between 2 groups (experimental groups A and B or experimental group and control group. For example, evaluate the experimental effectiveness of population education in 2 experimental and control groups by results posttest results) to see if the difference is statistically significant. The following conditions are required:

- 4/5 the product of the margin series must be greater than 5N (> 5N);
- Use qualitative measurements (bad-good; good-fair-average-weak-poor).

Above are just some common statistical math tests. In addition to the application conditions of the test, when used one should pay attention to the application situation and the overall purpose of the evaluation. Statistical math has advantages, but it is only a researcher's tool. From the data we have, how we classify depends on the expertise of the researcher. Moving from one metric to another also means moving to a different metric field so it will be different. The drawn conclusions should be checked.

3. Applying statistical probability in practice

All college students, whether math majors or students in economics, foreign trade, or even medicine, we all "have to" learn a subject that is: probability math. and statistics. Some practical applications of Statistics in business can be mentioned are: predictive analysis and data analysis model to make strategic suggestions, assist managers to make decisions in The moment is important and has great significance for the company's production and business activities. Statistical methods are the foundation in the data analysis process to determine future needs, trends in customer behavior and shopping habits.

Figure1. Business intelligence

Identify your target audience

Statistical research helps managers determine which customers are the target customers. By understanding customer information, consumer trends, purchasing power and preferences, business managers decide to develop products that better meet their customers' needs. To better understand what types of products and consumers need, how they will use them, companies need business analysts to understand and analyze the data properly. Fortunately, along with the development of data science, statistical software was born to help engineers analyze data effectively and quickly.

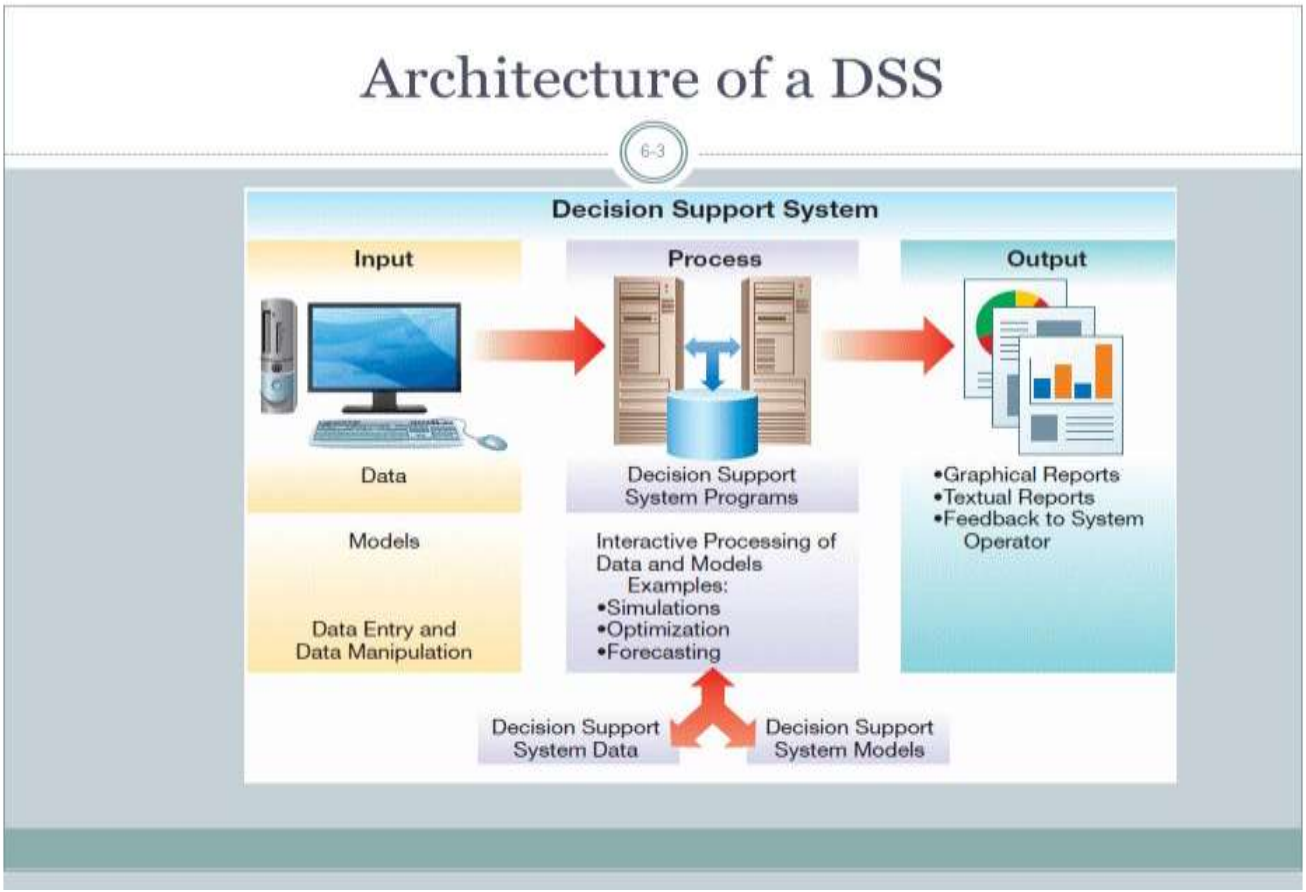

Figure 2. Decision support system

Promotional products

Statistical research is also used to decide on brands and to advertise products or services. From identifying and describing the current customer base, data analytics provide information about your marketing strategy, goals, which products will fit in which business channel, and when point for each customer segment. All of this information can be very useful for managers when making decisions about what types of messages to use and which products to include in advertising. In addition, statistical studies of media, groups of customers using a certain media, can help managers decide on where to buy advertising. To do that, business analytics engineers have to process enormous amounts of data; It is not only data collected about customers through transactions, but also data about communication providers. Strongly developed statistical tools together with techniques, algorithms in data analysis, statistical software, open source code help managers to quantify their decisions.

Decide on price

One of the key applications where statistical research is used in business is in making price decisions. Statistical analysis can help managers determine price trends, the consumer's sensitivity to higher or lower prices, and the ratio of production costs to prices. In addition to traditional methods such as regression models, time series models to determine the factors that affect product prices, price research models are used a lot in capital

investment projects. is valid, and is often referred to as "option pricing, the real options revolution". It can be seen that the development from theory to practice of statistics and data analysis is extremely important in the operation and development of a business.

4. Conclusion

In academic science research, qualitative analysis must be essential. But that does not mean biased towards arguments and disregard for data. Not excluding the possibility of using a table of figures with a few lines of notes and explanations enough to replace, sometimes even more convincing than written pages. In the process of training and maturing on pedagogical research, researchers need to try to make better use of the dialectical agreement between qualitative and quantitative analysis, towards the ability to capture the numbers instead. It can be affirmed that the presentation of data tables, graphs, models ... and how to use them to serve qualitative analysis is a reliable foundation for correct judgment about competence and qualifications. proficiency of a researcher.

Today, the strong development of information technology has been supporting a lot for researchers in the field of education. The application of statistical models to quantitative research in education and the use of specialized software regularly updated in terms of feature strength is becoming more and more popular. Therefore, this is also an urgent requirement for those who conduct educational scientific research if they want to improve the quality of research works. Currently, mathematical statistics are not only used in the processing and analysis of practical data, but also in other important stages of an educational research project, such as: Determination of sample size, construction tool, check for errors during implementation. Because each step in the research process has potential errors, which can damage or reduce the scientific value of the research.

There are many other applications in the business that managers need that, creating a connection between products, ensuring product quality as well as managing employee performance. But first, to have such meaningful analysis, businesses need to be conscious of data collection and management. Data is collected from many sources, within the production activities of the business, can also be data from the media, or data from surveys of customers. Owning a large and valuable data is an advantage in business if the manager knows how to access and use the data properly. With the strong development of big data, data science, data analysis models and software will be a powerful tool to help managers achieve high efficiency in business decision making.

References

i. *Robert J. Marzano, Debra J. Tickering - Jane E. Pollock (2011). Effective teaching methods. Vietnam Education Publishing House. Hanoi. (Translation by Nguyễn Hồng Vân).*
ii. *Robert J. Marzano (2011). Teaching arts and science. Vietnam Education Publishing House. (Translation by Nguyen Huu Chau).*
iii. *Lam Quang Thiep (2011). Measurement in education. Theory and application of the National University of Hanoi National University.*
iv. *Tran Trong Thuy (1992). Psychological science of diagnosis. Education Publishing House. Hanoi.*