Exactitude Of Evaluating Teachers' Performance Based On Students' Test Scores

Nguyen Van Phuong
Institute of Languages and Social Sciences
Ho Chi Minh City University of Transport
Ho Chi Minh City, Vietnam
phuonggtvt2007@gmail.com

***Abstract*-This research aimed at investigating the exactitude of evaluating teachers' performance only based on the students' results after the final tests at the University of Transport in Ho Chi Minh City. This educational managers' evaluation seems to create some doubt about both the validity and the reliability because students' examination results were probably affected by some physical and mental factors. A scientific research on the variety of the factors affecting on the students' results should be done seriously. The author of this research randomly selected 120 students in three classes of different class-sizes and of different majors as the participants. These students did the several tests which were designed by the lecturers in the English Department in an interval of three weeks without being informed the test's time. There were some changes of the number of exam-takers sharing the same desk, of supervisors and of test markers. Ten teachers of the English Department also participated as cooperators in this project. These teachers worked as the supervisors and also the marker of the students' papers. Some questionnaires and interviews were also designed to get further information about the test from both the students and the teachers. The findings of this research may serve as reference data for the administrators.**

Keywords-Evaluation, fairness, performance, result, test scores.

I. INTRODUCTION

The science of Language Testing not only helps us to build good tests but also serves as a tool to analyze test scores to draw sound conclusions on aspects related to the teaching and learning effectiveness. This research was conducted to investigate the accuracy of the evaluation of the teacher's performance based on the student's test scores at the Ho Chi Minh City University of Transport (UTH).

A. Rationale

Test scores, especially in the case of written tests, do not always reflect the true ability of test takers who might be influenced somehow by mental and physical factors while answering test. Thus, there arises some doubt about whether students' results in some exams can perfectly measure the teachers' performance.

B. Statement of Purpose

After having taught English at UTH since 2008, the researcher observed some students in his classes and recognized that some students who did not learn well in class still got high scores at some exams. On the contrary, the others who learnt actively and got high marks in class failed to get high scores in some exams. This research was conducted to find out whether the use of students' results in exams to evaluate teachers' teaching effectiveness at the UTH was accurate or not.

C. Research Questions

This research discusses these questions

- What are the contributors to test scores?
- To what extent can these test scores measure teachers' performance?
- Should test scores be considered the decisive criterion to evaluate the whole process of teaching?

D. Significance

The study aimed at proving the criteria which the administrators evaluate teachers' performance just based on the students' test scores seemed to be unconvincing and unreliable. Teachers' teaching ability should be measured by a wide range of valid information apart from students' test scores.

II. LITERATURE REVIEW

As far as we are concerned, validity and reliability are the two key issues in the field of language testing in second language performance. We would relate some aspects of reliability and some authors' ideas about this.

Reliability is, in fact, a prerequisite to validity in performance assessment in the sense that the test must provide consistent, replicable information

about candidates' language performance (Clark, 1975). That is, no test can achieve its tended purpose if the test results are unreliable. Since the administration of performance tests may vary in different contexts at different times, it may result in inconsistent ratings for the same examinee on different performance tests. We, therefore, should pay much attention to inter-examiners and intra-examiner reliability, which concerns consistency in eliciting test performance from the testee (Jones, 1979).

Hughes (2003) also shares the idea that the same student taking the same test but at a different time would not obtain the same score even other factors such as administration procedures, scoring procedures, environment are excellent. He suggests that teachers should administer and score test in the best effective method to make a similar score contributing much reliability to the test. To make tests more reliable, Hughes (2003) is concerned with two components of test reliability: the performance of candidates and the reliability of the scoring, the success of which requires us to

- Take enough samples of behavior;
- Do not allow candidates too much freedom;
- Write unambiguous items;
- Provide clear and explicit instructions;
- Ensure that tests are well laid out and perfectly legible;
- Be sure candidates familiar with format and testing techniques;
- Provide uniform and non-distracting conditions of administration;
- Use items that permit scoring which is as objective as possible;
- Make comparisons between candidates as direct as possible;
- Provide a detailed scoring key;
- Train scorers;
- Agree with acceptable responses and appropriate scores at outset of scoring;
- Identify candidates by number, not name;
- Employ multiple, independent scoring.

In addition, performance tests require human or mechanical raters' judgments. The reliability issue is generally more complicated when tests involve human raters because human judgments involve subjective interpretation on the part of the rater and may thus lead to disagreement (Mc Namara, 1996).

In a much more specific observation, Brown, J.D. (2005) makes a list of potential sources of error variance (or measurement error). According to Brown, a candidate's ability is not the whole thing deciding his result but there exist a number of indirect factors as follows:

- Variance due to environments;
- Variance due to administration procedures;
- Variance due to scoring procedures;
- Variance attributable to test and test items;
- Variance attributable to examinees;

The first environmental factor closely affects the performance of the students is the location of the test administration such as a library, a hall, a classroom or an auditorium. Obviously, different conditions of these places lead to different test results. The factors also include space, noise, ventilation, lighting, weather and so forth.

Secondly, some administration procedures having a considerable influence on students' performance must be the speed or clarification of directions, the quality of equipment and timing, the mechanics of testing or they may be the attitudes, the helpfulness or the anxiety level of the proctors.

The next potential source of error variance involves scoring procedures which mention human errors in doing the scoring, the subjective nature of the scoring procedures, evaluator inconsistencies or biases, and idiosyncrasies. The test and test items might result in some unexpected things for candidates in the way of test booklet clarity, answer sheet format, particular sample of items, items types, number of items, item quality and test security. The last immediate source of measurement error we cannot ignore is themselves examinees. Differences in physical characteristics, psychological factors and experiences among candidates do affect their performances (Brown, 2005). In short, all the authors we mentioned above observe and judge a candidate's result in various aspects attempting to give a clue that teachers should not be assessed to be good or bad just basing on their students' single final test results since there exist so many other influential factors to students' achievements. In this paper, we would try to clarify as well as prove that

administrator's policy in UTH is wrong or right in the language testing point of view.

III. METHODOLOGY

The researcher's purposes are to find out factors that both directly and indirectly influence students' test results and to prove that it seems to be unreasonable if teachers' teaching ability is evaluated through their students' test results. To achieve these purposes, we employed a number of methods including *(1) presenting the research questions, (2) choosing the participants, the instruments of the study, (3) selecting data collection procedures.*

A. Presenting Research Questions

When result of the first test (Test A) was reported, we found it strange because most of students who didn't study really well in the class got surprisingly high marks while a few good students only got average marks. We wanted to find out factors leading to this questionable result. We carried out a study to find out what these factors were. To achieve this purpose, we asked students of three classes CN07N3, HH07A, and MT07A to do test A again three weeks after the first test. We thought that three weeks was long enough for them to forget the content of that test but it was also so short that during that time they could not get any more knowledge to do the test better. The reasons why we chose these classes were (1) that they were of different class sizes, (2) that they were of different levels and (3) they were taught by different teachers who were of different ages. Two weeks before answering test B, students were invited to answer a questionnaire and simultaneously ten out of twenty teachers of the English department were invited to answer the researcher's interview questions. The reason why we selected these teachers was they were in charge of these classes or others of the same level. The results we got from these above activities could help us work out some suggestions on how to evaluate teacher's teaching effectiveness based on the result of a final test objectively and reliably. In order to reach an accurate objective, the researcher posed a number of guiding questions for the study.

- The fairness issue: Did all the invigilators maximize objectivity to give each student an equal chance to do the test well?
- The test construction: Is the time allowed rational compared with the length of the test? Does the difficulty of the test suit students' level? Are the instructions and content of the test clear enough for students to avoid misunderstanding?
- The test administration: Are students often nervous during a test? Are teachers under pressure because of the long test? Do the equipment and classroom meet the requirement of administration? Are all the necessary physical conditions for the test met?
- The test scoring: Are there any scoring mistakes?

In order to answer all of these questions, a research design with appropriate components is indispensable.

B. Research Design

1) Participants

The study was carried out with the participating of two following groups:

Students: The first group of the participants were 120 voluntary students from the three pre-intermediate general English classes at UTH marked as HH07A (58 students), MT07A (42 students), CN07N3 (20 students). All of them were randomly invited to participate in the study because they are of the same level.

Teachers: The second group of the subjects included 10 members of the teaching staff who are in charge of teaching various general English classes at UTH. Three male teachers and seven female teachers aging from 28 to 40 with M.A. degree in TESOL were invited to answer the researcher's interview questions.

2) Instruments

Data was collected by using three different types of instruments including the students' questionnaire, the test results and the interviewing teachers.

a) The students' questionnaire

The students' 16-item questionnaire was divided into two major parts: In the first part of the questionnaire, the subjects were required to provide some of their background information on their name, age, gender and class. These details helped to establish the profiles of the students participating in the study and helped him make recommendations for Administrator Board to evaluate the teaching staff at

UTH. The second part, which was also the main section of the questionnaire, aimed at eliciting information on factors affecting students' result of the two tests. This part can be summarized as follows:

Questions 6-9: The aim of these questions was to help us get information about the ease of the test. Students may lose their self-possession when they realize that the test is so long or so difficult while the time allowed is so short. Unavoidably, that will have a bad effect on students' psychology, and thus, negatively influence the test result.

Question 10: This question was included in the questionnaire because we wanted to know whether the instructions of the test were clear and understandable for students.

Question 11: This question was asked to find out about student density in the testing room. It is quite possible for students to copy each other's papers if the testing room is densely populated. In such cases the test result is not reliable.

Questions 12-14: Through these questions, information about surrounding environments and about students' physical as well as mental conditions was collected. These factors, as we believe, also have a little influence on the test result.

Question 15: This question was designed to help us know about students' attitude towards the test. If they are serious about the test, then they will have good preparation for it, and thus the result will certainly be satisfying, and vice versa.

Question 16: The purpose of this question was to find out whether the invigilators worked seriously or not. Clearly invigilators' working attitude has a great impact on the test result.

b) The interviews with teachers

To get teachers' opinions on a wide of various matters influencing on students' result and the way to evaluate teachers' effectiveness by the Administrator, he carried out an interview with the participation of ten teachers who were assigned to teach pre-intermediate English classes at UTH. A list of seven questions for interview was prepared beforehand so that the following points from the teachers' viewpoint would be collected and then taken into consideration:

Questions 1 and 2: The purpose of these questions was to get teachers' comments on the test items.

Question 3: This question was asked to help us know whether students sat too close to each other in the room.

Question 4: Through this question, we got some ideas about teachers' working attitude during the time when they invigilated the test.

Question 5: This question was designed to help us know how the marking was carried out.

Question 6: The aim of this question was to get teachers' opinion about the policy issued by the Administrator.

Question 7: This question was aimed to invite teachers to give their suggested solutions to the only use of students' result to evaluate teachers' performance.

C. Collection Procedures

The questionnaire was delivered to the students in their regular classroom at their convenient time. They were informed of the purpose of the study and of their significant contribution as the subjects to the success of the study.

All the questions in the questionnaire were explained clearly so that nobody misunderstood. Students answered the questionnaires and teachers collected them on the spot. The semi-structured interview consisted of seven questions prepared in advance. Each of the teacher subjects was interviewed personally and in a relaxed manner.

All the answers collected from the interviewees were recorded so that not any information in their answers would be lost or misinterpreted.

IV. DATA ANALYSIS AND FINDINGS

A. Learners' questionnaire

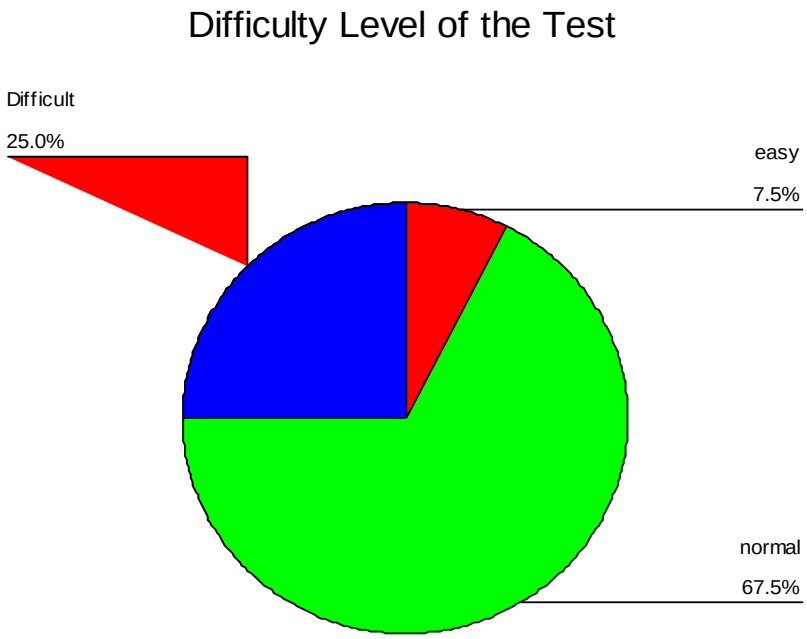

Figure 1. Question 6:
The difficulty of the test.

As much as 67.5% of the students who were asked about the difficulty of the test thought that the test was not so difficult for them. Only 25% of them claimed that the test was too difficult. 7.5% of them found the test rather easy.

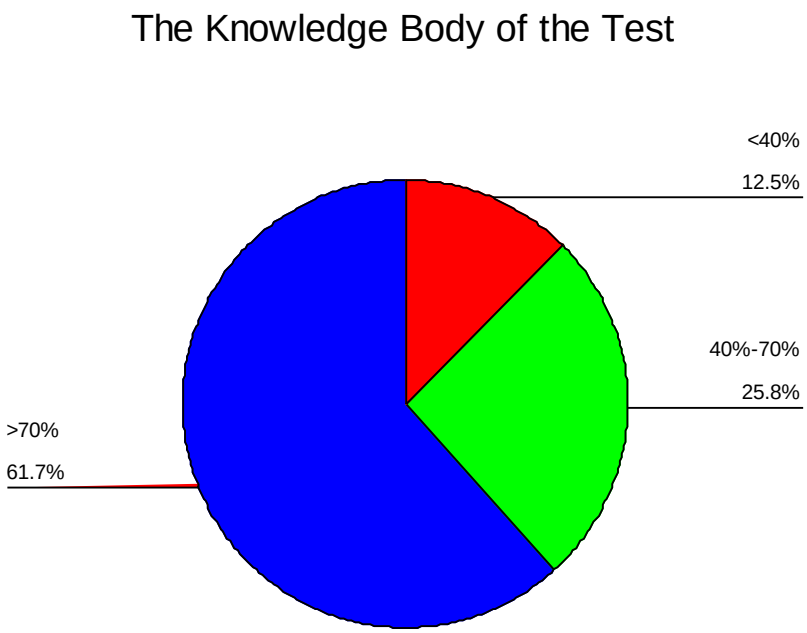

Figure 2. Question 7:
The knowledge of the test.

Over 60% students said that over 70% of the knowledge tested was taught to them in class whereas about 25% thought that amount of knowledge was only 40%-70% and about 13% students thought that percentage was less than 40%.

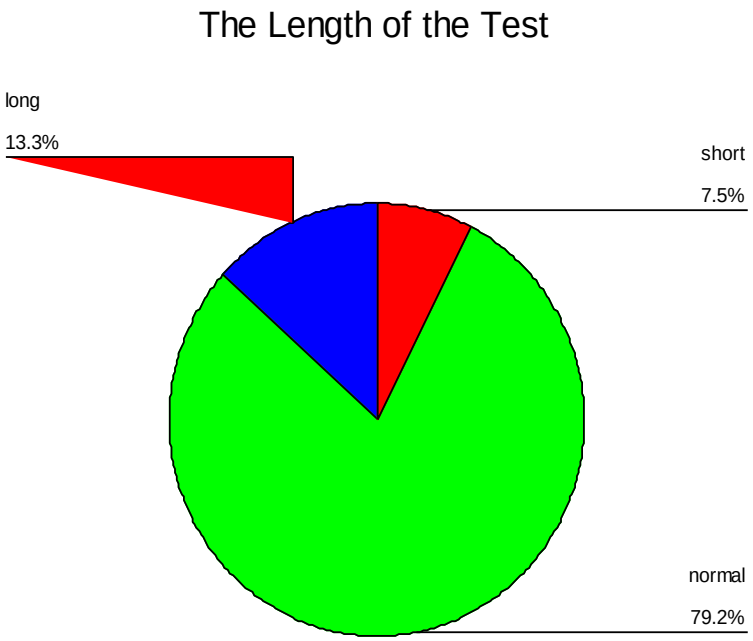

Figure 3. Question 8: The length of the test.

When asked about the length of the test, about 80% of the students said that the test was of medium length while about 13% of them thought that the test was too long and the other 8% said that it was too short.

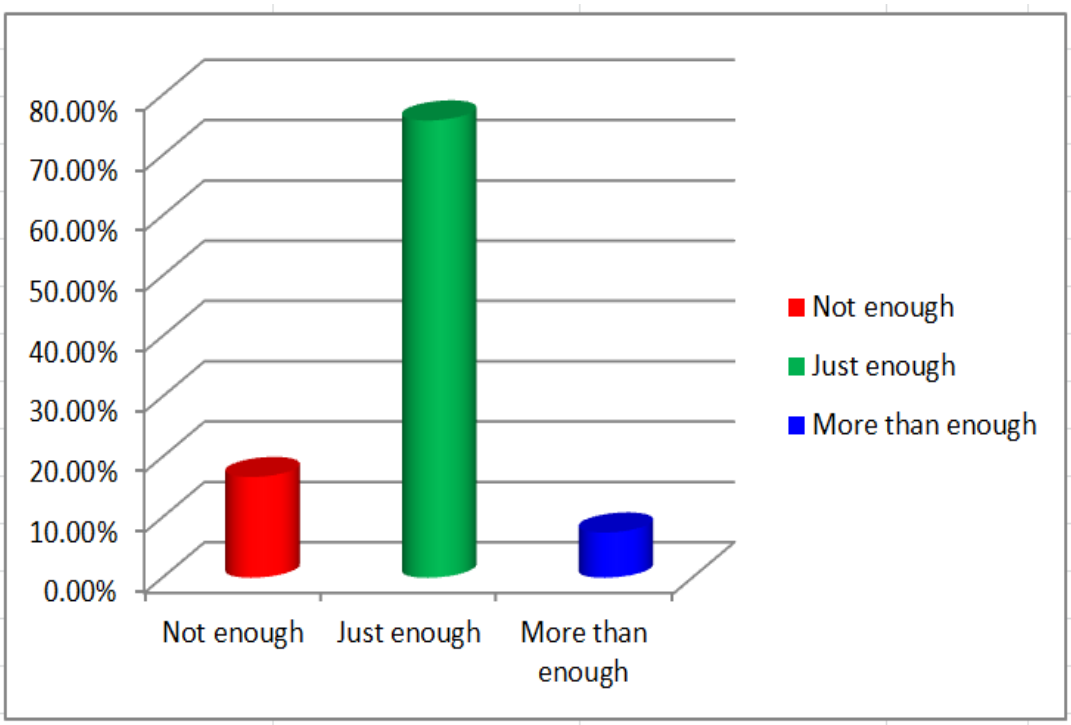

Figure 4. Question 9: The times allotted to answer the test.

About 76% of the students thought that the time length is enough. 17% said that the time allowed was too short for them to cover all the tests whereas 7.5% thought that the time they really needed to do the test was not that much.

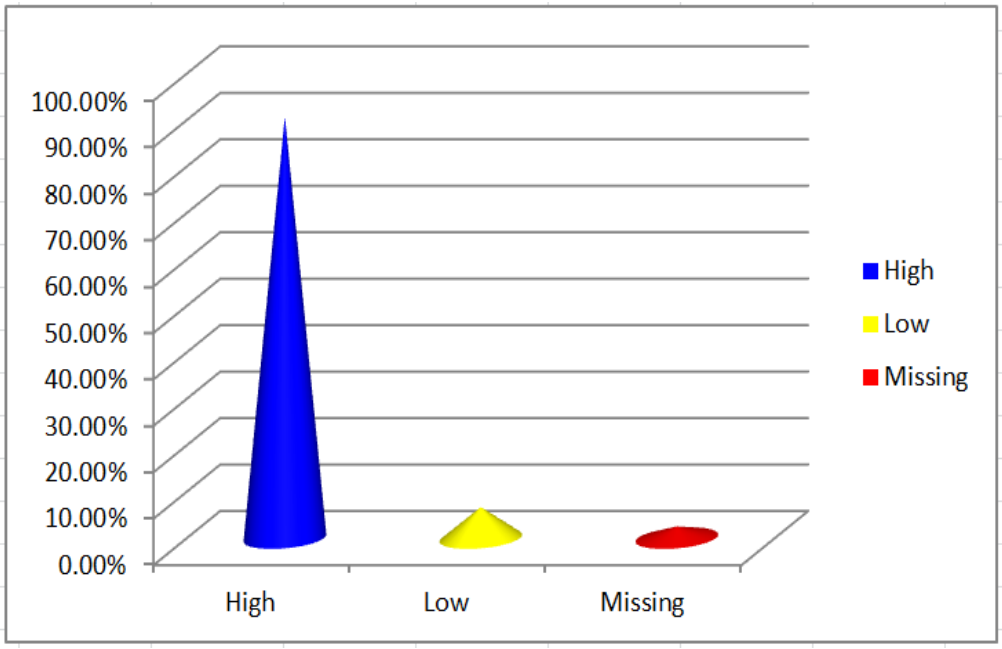

Figure 5. Question 10: The clarity of the test.

Only 97.5% of the students who were asked about the clarity of the test gave answer, 93.2% of whom claimed that the test instructions were clear and understandable enough for them. When asked about the position sitting in the testing rooms, 74.2% of the examinees said that there were three students sitting in the same table. The density might lead to the students' copying. About the mental and physical conditions of the students as well as the surrounding environment while doing the test, 96.7% of the exam-takers agreed that they satisfied with these conditions.

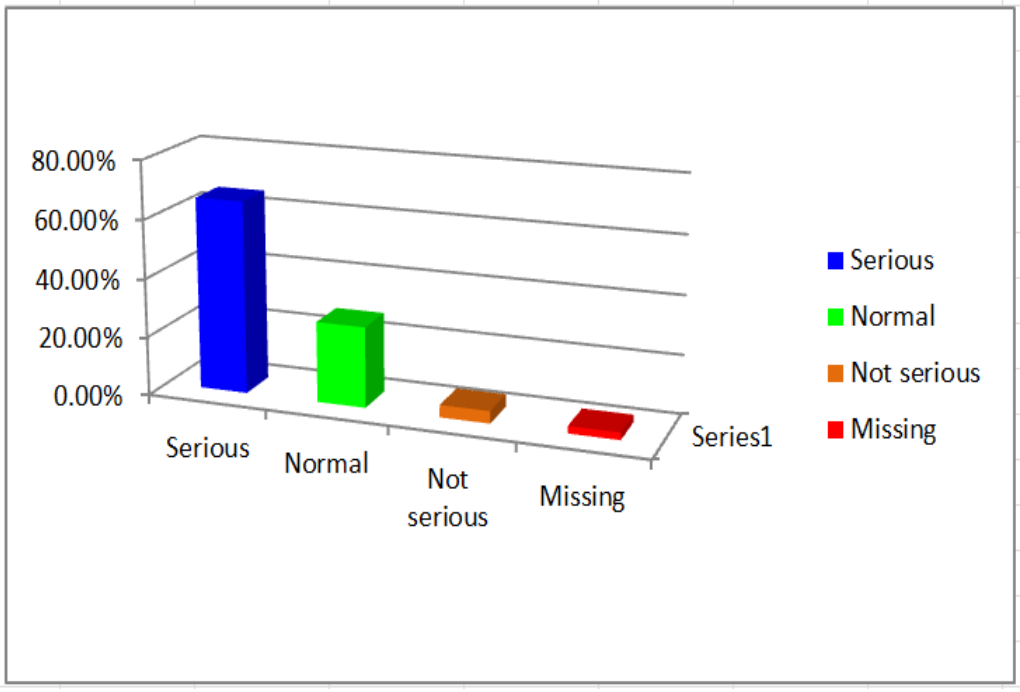

Figure 6. Question 16: The attitude of the examiners.

In order to find out the examiners' attitude while they organized the tests, 65.8% of the students answered the examiners were serious and 27.5% of the teachers kept normal attitude. Thus, it is considered that the examiners were responsible for their duty.

B. Interview teachers

All of the ten teachers asked said that the test was suitable for students' level because 90% of the knowledge tested was from the textbook. The other 10% was for excellent students.100% of the teachers claimed that the allowed time to

answer the test was reasonable because they were all asked to design a 60-minute test. They all agreed that it was ideal to arrange one student per table but in cases when there were not enough seats as required due to the large class size, they had to arrange two and even three students per table. 08 out of 10 teachers interviewed thought that examiners should be serious during the test administration while the other two hold the viewpoint that we teachers should not be so strict in the exam room because that will cause unnecessary stress to students. All the interviewees asserted that it was obviously unfair and unreasonable to base the evaluation of teachers' teaching ability on students' test results only. To be fair, in their opinion, we should be serious in our testing process, not only in a test designing a stage but in test administration stage and in test scoring stage as well. How to keep a reasonable student density in the exam room was another important issue we should put into careful consideration.

C. Collecting test B result and data analysis

When comparing the results of the test A and test B, it was recognized that the scores the students of the class CN07N3 got did not change much whereas there were a remarkable change in the results of the students in the class HH07A and MT07A. The following is the detail.

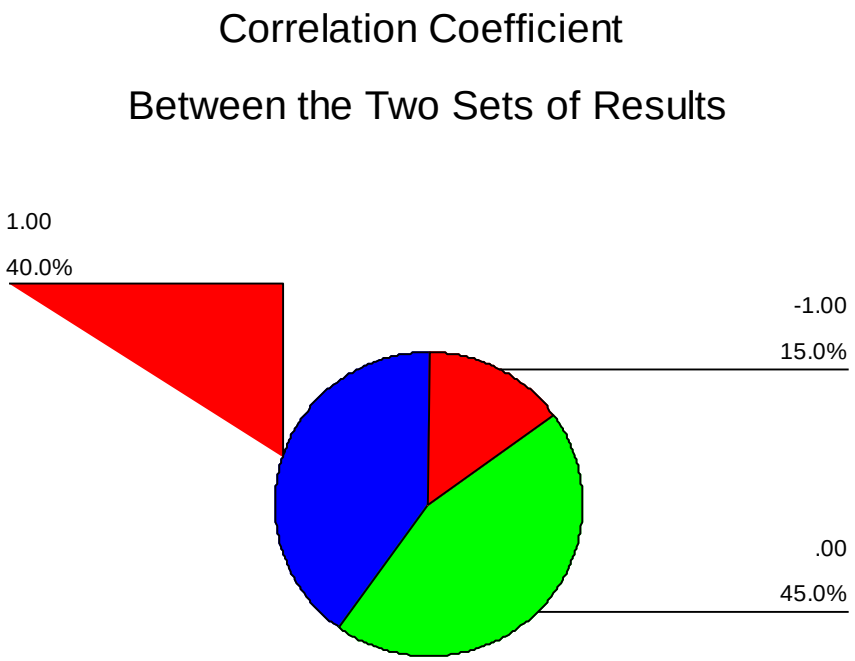

Figure 7. The comparison of three classes' result.

As seen from the table and the chart above, the difference in results between the two tests varied from -1 to +1. The number of students who got the same marks for the two tests was 09, comprising up to 45% of the total number of the students of the class CN07N3 and around that percentage of students got the difference of +1.

Only 03 students, making up 15%, got the difference of -1. In general, the students' results of the two tests didn't change much.

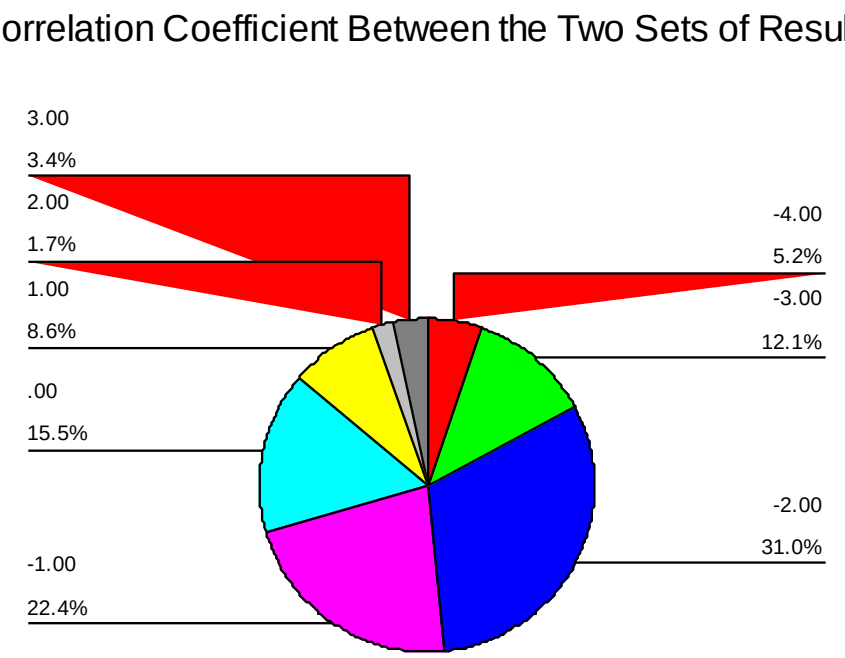

Figure 8. The comparison of two tests' result.

Looking at the table and the chart above, we could easily recognize that there was a remarkable change in the results of the two tests. As much as 70.7% of the students in the class got the marks which were from 01 to 04 marks lower in the second test while only 13.8% of the students received better marks. The results of 15.5% of the students stayed the same in the two times of testing. In short, the majority of the students got different marks in the second test.

V. CONCLUSIONS AND RECOMMENDATIONS

From the above-listed outcomes, it was noticed that there were a wide range of factors influencing on students' test results. First, the arrangement of the seats for the test takers and the examiners' attitude might either prevent or foster plagiarism among students. Second, the students' physical and mental condition might affect their test scores. Third, the environment conditions such as weather or noise could distract the test takers from their concentration on the test. As a result of these, students might get lower marks or higher ones compared with their real ability. Thus, that the administrators evaluated teachers' performance just based on students' test scores was likely totally unconvincing. The administrators should consider more different valid information channels to make a right decision on teachers' performance such as participating some teaching periods to observe the teachers' teaching practices, following the teachers' self-training, teachers' qualifications, their personal characteristics, teaching experience, the ability to motivate interactions with students and the ability to create a positive classroom environment.

In case of using students' test scores to measure the teachers' teaching ability, the administrators should consider students' result as one of the criteria to evaluate teachers' performance. As stated by Goe (2008), test scores are limited in the information they can provide. The student learning cannot reasonably be attributed to the activities of just one teacher – it may be influenced by other teachers, peers, study resources, school climate and family.

REFERENCES

[1] J. L. D. Clark, "Foreign Language Testing: Theory and Practice. Language and the Teacher: A Series in Applied Linguistics," vol. 15, Philadelphia, USA: Center for Curriculum Development, Inc., 1972. Available:https://eric.ed.gov/?id=ED060771.Accessed on: 14/06/2022.

[2] L. R. Jones and S. Bernard, "Testing Language Proficiency," Washington, DC, USA: Center for Applied Linguistics. Available: https://eric.ed.gov/?id=ED107161. Accessed on: 14/06/2022.

[3] A. Hughes, "Testing for Language Teachers," Second Edition, Cambridge, UK: Cambridge University Press, 2003.

[4] T.F. McNamara, “Measuring Second Language Performance,” First Edition, Essex, London: Addison Wesley Longman Ltd., 1996.

[5] J. D. Brown, “Testing in Language Programs,” Second Edition, Singapore: The McGraw-Hill Companies, Inc, 2005.

[6] L. Goe, C. Bell and O. Little, “Approaches to evaluating teacher effectiveness: A research synthesis, Washington, DC, USA: National Comprehensive Center for Teacher Quality, 2008. Available: https://gtlcenter.org/sites/default/files/docs/EvaluatingTeachEffectiveness.pdf. Accessed on: 14/06/2022.