│Volume 7 │Number 2 │October 2022│E-ISSN: 2503-4405│P-ISSN: 2580-3441│

English Summative Tests: The Quality of Its Items

Thresia Trivict Semiun[1], Maria Wihelmina Wisrance[2], Merlin Helentina Napitupulu[3]

[1, 2, 3]Universitas Timor, Indonesia

[1] semiunthresia@gmail.com, [2]mariawihelminaw@gmail.com, [3]merlinn10@gmail.com

Abstract

It is crucial to implement evaluation after the teaching and learning process. Evaluation will reflect the success of teaching and more important the achievement of the students. Therefore, EFL teachers should develop a good test to measure students' achievement. This study analyzed multiple-choice items of English summative tests constructed by junior high school EFL teachers in Kupang, NTT. The result of this analysis functions as feedback to the English teachers on the quality of English summative tests they had created. This research was descriptive research with documentation for data collection.The English summative tests for grades VIII and IX were collected and then analyzed by using ITEMAN software to reveal item difficulty, item discrimination, and distracters effectiveness of the tests. The findings revealed that the English summative tests were developed with easy items. However, the tests still had a good discriminatory level. The test items which had all distracters perform well were only half of the total items.

Keywords: *English summative tests, the quality, items*

INTRODUCTION

Evaluation refers to a systematic process to determine which instructional goals are achieved by students (Gronlund, 1982). Evaluating educational outcomes is one of the responsibilities of teachers. The accomplishment of instructional objectives is done through a measurement. By measuring and evaluating, teachers can diagnose the strengths and weaknesses of their students for then take action so that progress and improvement can be done (Hadi & Kusumawati, 2018).According to Allen (2004), teachers examine learning in the day-to-day classroom and evaluate students' attainment in the entire curriculum through assessment systematically. Further, teacher applied evaluation to refine teaching-learning instructions to improve students' learning. Evaluation, though, should be implemented, to promote student learning, not because the educational institutions require it.

Evaluation involves the use of empirical data on students' achievement obtained after administering a test. Tests are essential to determine the success of teaching and learning process and to measure the achievement of the students (Wisrance & Semiun, 2020).In evaluating the students, the achievement test seems to suit very well. Fraenkel & Wallen (2006) assert that achievement tests play a prominent role and are a widely used method to assess student achievement in classroom instruction. It aims to measure the mastery of a set of learning objectives. It also reflects the efficiency of a school program or curriculum in contributing to student learning. Koretz (2002) argues that scores on most achievement tests are not truly meaningful or informative. They fall into two senses. One of which is in the traditional statistical sense where scores include measurement errors because they are venerable to inflation. The lack of cheating controls when the students do the test is an example. When students obtain higher scores through cheating, the test is neither reliable nor valid.

In assessment, there are two objectives which are for helping to learn and summarizing the learning. The formative assessment is applied to maintain teaching and learning. Meanwhile, the use of summative assessment is for recording and reporting (Allen, 2004). The summative test has a greater emphasis on students' achievement. The result of a summative test will be employed to assign grades to students. The summative test involves collecting evidences about students' achievement in a systematic way to be reported at a particular time based on teachers' professional judgment (Harlen, 2004).

Most of the achievement tests are teacher-made. Teacher-made tests are those which have not been standardized. According to Djiwandono (2011), based on the method of preparation and development, tests can be distinguished between standardized tests and teacher-made tests. Standardized tests are compiled and developed based on strict guidelines, requirements, and procedures to produce tests that have good test characteristics and that have been reviewed as planned. Meanwhile, teachers prepare and develop teacher-made tests as part of daily teaching tasks to evaluate the implementation of teaching, including the students' learning progress. Arikunto (2010) states that teacher-made test has different validity and reliability compared to standardized test. Brown (2004) explains that teacher-made tests are composed to be easy and quick to administer and score, less expensive, and do not require extensively trained test givers as teachers have enough competence to develop a good test. Teacher-made tests have the advantage of being directly related to the content taught in the individual classroom. The tests should follow a detailed course syllabus, books, and other materials used (Rudner & Schafer, 2002).

The information yielded by the educational assessment massively affects students' learning, education instruction, and curriculum. According to Bachman & Palmer (1996), language tests are valuable tools for providing relevant information about language teaching. It provides information on the effectiveness of the teaching programs and information on whether or not the students are ready to move on to another unit of instruction or assign grades based on students' achievements. The tests are tools for evaluating the relevance of instructional objectives and instructional materials with specific activities designed in the lesson plan.

According to Bachman & Palmer (1996), the educational assessment should be of good quality and contain a minimum number of errors. Therefore, the teachers need to conduct a test analysis to know the quality of the test. If the test contains many errors, it is possible that the evaluation already goes wrong and is unfair. However, teachers rarely perform a test analysis due to the following reasons. First, teachers do not understand or ignore the importance of a valid and reliable test. Second, teachers do not know the method to analyze tests, and third, teachers probably feel that test analysis is time-consuming (Santyasa, 2005). Nana *et al.* (2018) asserted that teachers lack techniques and skills in constructing the test. Teachers have close daily contact with the whole testing process. Maharani, et. al (2020) revealed that the English teachers didn't conduct item analysis. The teacher just distributed the tests to the students. "An English teacher who constructed a test to assess students' learning outcomes had never conducted the item analysis." (Karim, *et al.*, 2021)

Muhson*et al*. (2017) noted that "item analysis program is designed to measure student achievement and instructional effectiveness. It is a valuable way to improve the test items for later usage tests. After conducting item analysis, the test developers could eliminate ambiguous or misleading items of the tests. Item analysis also can help the test makers to increase their skills in creating the test. It helps identify questions that are not performing well and also specific subject content that needs greater emphasis or clarity. Item difficulty, item discrimination, and the performance of distracters should be considered (Wells &Wollack, 2003).

Item difficulty is a measure of the proportion of test-takers who have answered an item correctly. It refers to the p-value (Kohoe in Mahroof& Saeed, 2021). In practical terms, item difficulty range from 0.00-1.00. Good questions should have moderate difficulty value because very easy level questions may not sufficiently challenge the ablest students. On the other way, very difficult level questions may produce frustration among students.Item discrimination (ID or D) provides information about students who know the material well and vice versa. Item difficulty and item discrimination are associated with one another. If the item difficulty of the test is very easy or difficult, the test item will have little discrimination. Conversely, if the test items are of moderate difficulty, the items will be more discriminating. DiBattista & Kurzawa (2011) asserted that "the discriminatory power of a multiple-choice item depends heavily on the quality of its distracters." According to DiBattista & Kurzawa (2011) "distracters are the multiple-choice options that are plausible but not the correct answer". An effective distracter looks plausible to less knowledgeable students and lures them away from keyed options. Yet, it does not lure students who are well-informed about the topic being tested. Haladyna & Downing (1993) claim that at least 5% of test-takers should select each item distracters. In addition, distracters also should be able to discriminate between stronger and weaker test takers. Thus, test takers who possess higher overall scores must select the keyed option more often than those with lower scores (Djiwandono, 2011).

Previous research focused on items of English tests has been conducted by scholars. Salija*et al*.(2018) found that the English test was constructed by easy and medium items, and only 50% of distractors functioned. Pradanti *et al*. (2018) explained that the English Summative test only had 46% of multiple-choice items fulfilling the criteria of good test items. Hence, the test failed to meet the quality of a good multiple-choice test. Studies have pointed out the difficulty and discriminatory power of multiple-choice items that change when dysfunctional distractors are either replaced or deleted (e.g., Cizek & O'Day, 1994). However, few have looked at the quality of distractors in multiple-choice tests specifically designed for classroom use. Tarrant *et al*. (2009) defined a functional distractor as one that was selected by at least 5% of examinees and was also negatively correlated with test scores. Tarrant et al. noted that they excluded tests with the reliability of less than 0.70 because if the lower reliability tests had been included, the percentage of functional distractors would likely have been even lower. In short, the research result suggests that many distractors used on classroom tests cannot function properly. Inferring from the previous research (Cizek & O'Day, 1994; Santyasa, 2005; Salija*et al*., 2018; Pradanti*et al*., 2018), item analysis is rarely conducted by teachers, and studies that analyzed functional items in multiple-choice classroom test, revealed that the test has low quality or below moderate. Since the tests are teacher-made, the quality of the tests tends to be different.

There were a lot of studies focused on item analysis of teacher-made tests. However, the study on the quality of an item of English summative tests is still rare in several districts located in East Nusa Tenggara Province. The present study was executed to analyze the quality of multiple-choice tests by referring to previously stated studies. The location was in Kupang, East Nusa Tenggara Province. The focus of this study is to analyze item difficulty, item discrimination, and item distracters of English Summative tests constructed by two different English teachers. Researchers expect that this study can be fruitful for improving the quality of English test items, particularly for some schools in Kupang as a qualified test can be a valid measurement media to evaluate the whole process of teaching and learning. It also helps us to know whether or not the instructional goals of the learning process are achieved (Gronlund, 1982). Furthermore, it is an important device to uphold test effectiveness and fairness. The English teachers could revise or remove the poor and not acceptable items if they want to use the test again. Through conducting item analysis, the quality of items can be improved Gronlund (1998). Then, this study also provides input to the school, where it needs to train the teachers to construct a good English summative test because the test is used to measure students' achievements at the end of the semester.

METHOD

This study was descriptive research to describe the quality of the items in terms of item difficulty, item discrimination, and distracter effectiveness. The tests used in this research were English summative tests to test grade VIII and IX students. The tests were constructed by two different English teachers who teach grade VIII and IX students. Therefore, there were two different packages of tests that tests students' achievement at the end of the semester. These tests were multiple-choice tests with five optional answers. The tests were aimed at measuring students' reading and writing skills, 40 items measured reading ability while 10 items measured writing ability. The selective school was a public junior high school that was well-known as a favorite school. The A accredited school has been acknowledged for the facilities and the quality of the teachers. The students' grades are above the standard. Whether the grades obtained by the students are in line with the difficulty and discrimination of the questions in the test or are the students smart. The assumption is the teachers can construct a good quality test since they work at *A* accredited school.

The quality of multiple-choice test items was examined by analyzing the responses that students made to each item. The analysis covered item validity, item difficulty (p), item discrimination (ID), and distracters effectiveness by using different formulas then run by ITEMAN 3.0 software. At the end of the item analysis report, the difficulty and discrimination index would be based on Brown and Arikunto as quoted by Salwa (2012). Meanwhile, the effectiveness of distracters was estimated by the number of students who answered the distracters. At least 5% of the students should choose them. If no student chose the distracters, it meant that the distracters could not perform well and that should be removed (Haladyna & Downing, as cited in DiBastista & Kurzawa, 2011).

RESULTS AND DISCUSSION

This section provides information related to item difficulty, item discrimination, and distracters effectiveness. The following are the summaries of the data analysis.

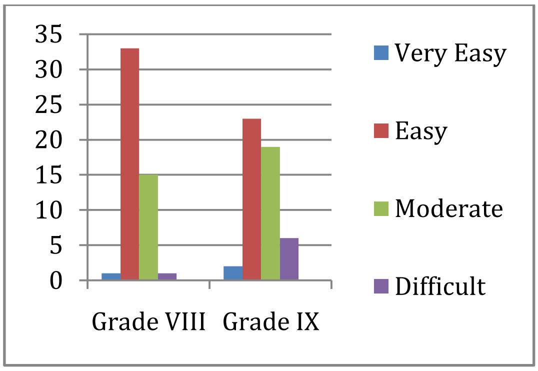

Figure 1. The Distribution of Classified Difficulty Index

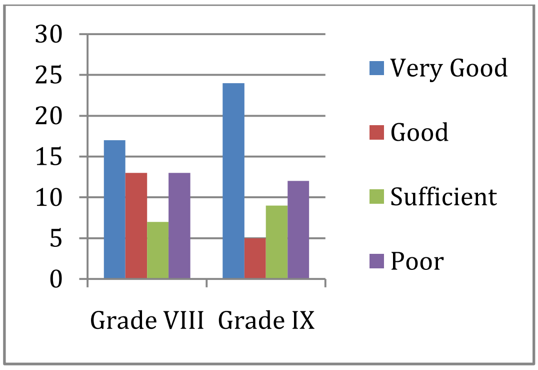

Figure 2. The Distribution of Classified Discrimination Index

Both of the tests are constructed mostly by easy items. However, the tests still have a very good discrimination level. There are pieces of evidence that the test is good tests perceived by the quality of its items. These items might look easy because the students have mastered the material well. If the items are easy and have a poor discrimination level, then it can be interpreted that the quality of the tests is not good. The result of the analysis indicates that difficult items could possess a poor and good discrimination index. These cases imply different views as portrayed in table 1.

Table 1. The excerpt of Item Difficulty (P) & Item Discrimination (D)

Item	Item Difficulty (P)	Item Discrimination (D)
43	0.277	0.557
48	0.128	-0.020

Item number 43 of the Grade VIII test indicates that the question is difficult but has very good discrimination power. This item is accepted. Meanwhile, item number 48 indicates that the question is difficult and the question is answered correctly by most of the weak students. Probably, students guessed the answer and most of them are not well-informed about the indicator tested in this item. Thus, the teachers need to greatly emphasize teaching students this material domain that aims to measure writing skills.

Concerning the analysis result, the minimum score of the tests is 21, and 23 out of 50, and the maximum score of the tests are 48out of 50 for grade VIII and IX. From the scores, it can be interpreted that among the students involved in the present study, there are weak students and good students academically. In addition, the median score of the tests are, 41, and 29 out of 50 for grades VIII, and IX respectively. It can be assumed that there are many good students of grade VIII involved in the present study. On the contrary, the grade IX test shows low median scores which indicates that of the students who participated in this study, there are good students as well as weak students academically. Hence, the result of item difficulty of grade IX indicates that there are 19 moderate items and 23 easy items, and 6 difficult items. Between the tests, the grade IX test is considered better compared to grade VIII tests of the item difficulty.

In this study, researchers consider the value of the difficulty index and discrimination index to decide whether an item can be accepted, revised, or removed. The test items are accepted if the value of the difficulty index and discrimination index is in balance. The decision is based on the index proposed by experts in the field of educational measurement (Brown, 2004; Arikunto, 2006 as quoted by Salwa, 2012). The total of items that can be reused is greater than those that should be revised or removed. The accepted items can be reused next time. However, the revised items should be replaced with better ones. The results of the items are presented in table 2.

Table 2. The Decision on the Test Items in the Tests

Decision	Grade VIII	Grade IX
Accepted	60% or 30 items	60% or 30 items
Revised/Removed	40% or 20 items	40% or 20 items

Besides the index of difficulty and discrimination, it is also important to check the effectiveness of distracters. A good distracter or a good alternative answer should have a distribution index of more than 0.025 or 2.5% found by using ITEMAN software. The summaries of analyzing distracters distribution are portrayed in the following figure.

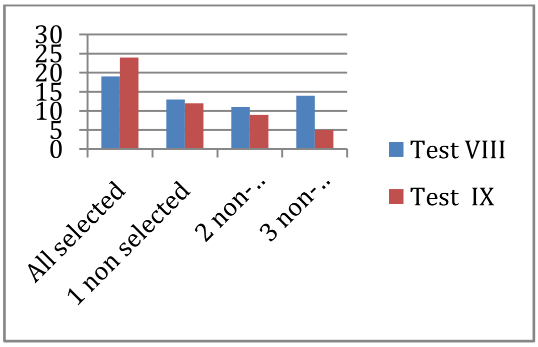

Figure 3. The Distribution of Distracters within the Tests

The figure shows information dealing with 50 items in the test which contains all selected distracters, 1 non-selected distracter, 2-non selective distracters, and 3-non selected distracters. First, the test total of all selected distracters is almost similar for two tests, i.e., in 19 and 24 test items respectively for the test for grades VIII and IX. Next, the total of 1 non-selected distracter is around the same number for tests, i.e., in 13 and 12 test items respectively for the test grade VIII and IX. Third, the test for grade IX contains the lower total of 2 non-selected distracters in 9 test items. Last, the test for grade VIII is higher with a total of 14 test items for 3 non-selected distracters. Therefore, within the test, the test items which have all distracters perform well are only half of the total 50 items or almost 25 test items. The ineffective distracters have to be modified to function more effectively when reused in the future. This method can help to maximize reliability because the greater the number of plausible distracters, the more accurate and reliable the test typically becomes (DiBatista & Kurzawa, 2011).

Some inferences can be drawn based on the findings. First, the tests are constructed with easy items. The current research is similar to the result conducted by Wiyasa *et al.* (2019). The items might look easy because the teacher has given the answers and the students remember them (Semiun & Luruk, 2020). In the test for grades VIII and IX, there are many items with good discrimination power. The items look easy because there are many good students involved in the study not because the items are below students' level of competence. An item with good discrimination power might be an important factor in the effectiveness of distractors (Maharani &Putro, 2020). For the distracters, the total of all functional distracters and two functional distracters in the item is greater than 1, and zero functional distracters in the item. When all the functional distracters and two functional distracters could function effectively, it might be assumed that the materials tested by the items are new material or never given to students during instruction. When writing a multiple choice test, item construction of the distractors is the major challenges (Malec &Krzeminska-Adamek, 2020). Therefore, the items that have a poor discrimination index need replacement or modification with the new one. This is similar to what was suggested by Toksöz & Ertunç (2017) Danuwijaya (2018) in their research findings.

Current research is limited to the English summative tests in the form of multiple-choice constructed by the English teachers to measure VIII and IX graders' achievement in reading and writing skills at the end of the semester. Therefore, the findings are only limited to that area. By investigating the item analysis, the teachers are informed and self-reflected on the quality of tests that are used to measure students' achievement at the end of the semester.

CONCLUSION

In conclusion, English summative tests of grades VIII and IX are constructed with good items and the items can function properly. In terms of difficulty level, the tests are mostly constructed with easy items. The items look easy presumably because the students are clever or because the same materials have already been given during teaching and learning instruction so that the students can remember the answers easily. For the discrimination index result, most of the items can discriminate between students who are good and weak. Next, in the case of distracters, it is concluded that most distracters or alternative answers can perform well in the test for grades VIII, and IX. On the contrary, there is a fewer number of poor distracters. There are some implications to address for teachers. The teachers need to conduct an item analysis every semester to check the performance of items. Item analysis should be done by teachers to replace very easy and very difficult items. In addition, when distracters are identified as being non-functional, teachers may remove and create a new distracter. Therefore, it is better if the school has a program to train the teachers in analyzing test items. This research is limited to the item analysis of English summative tests, so it is suggested to future researchers to analyze English formative tests (Mid-semester test) to obtain a border picture. The students also could be interviewed to obtain qualitative data regarding the difficulty of the test.

REFERENCES

Allen, M. (2004). *Assessing Academic Programs in Higher Education*. Bolton: Anker Publishing Company, Inc.

Arifin, Z. (2013). *Evaluasi Pembelajaran: Prinsip, Teknik, Prosedur*. Bandung: PT. Remaja Rosdakarya.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice. Designing and Developing Useful Language Tests.* New York: Oxford University Press.

Brown, H. D. (2004). *Language Assessment. Principles and Classroom Practices*. NewYork: Pearson Education, Inc.

Cizek, G., & O'Day, D. (1994) Further investigation of nonfunctioning options in multiple choice test items. *Educational and Psychological Measurement, 54,* 861-872.http://dx.doi.org/10.1177/0013164494054004002

Danuwijaya, A. A. (2018). Item analysis of reading comprehension test for postgraduate students. *English Review: Journal of English Education*, *7*(1), 29-40. doi: 10.25134/erjee.v7i1.1493.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests.*The Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2). https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Djiwandono, S. (2011).*Tes Bahasa. Pengangan bagi Pengajar Bahasa (*2ndₑd.*)*. Jakarta: PT. Indeks Jakarta.

Fraenkel, J.,& Wallen, N. (2006). *How to Design and Evaluate Research in Education* (6thed.). New York: McGraw-Hill.

Gronlund, N. E. (1982). *Measurement and Evaluating in Teaching*(4thed.). NewYork: Macmillan

Gronlund, N. E. (1998). *Assessment of Student Achievement*(6th ed.). Boston: Allyn and Bacon

Hadi, S.,& Kusumawati. (2018). An analysis of multiple-choice questions (MCQs): item and test statistics from Mathematics assessments in senior high school. *Research and Evaluation in Education,* 4(1), 70-78.https://doi.org/10.21831/reid.v4i1.20202

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*(4), 999 1010. https://doi.org/10.1177/0013164493053004013

Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature, and Culture*, *6* (2), 256-269. http://dx.doi.org/10.30659/e.6.2.256-269

Koretz, D. (2002). Limitation in the Use of Achievement Tests as Measures of Educators Productivity. *The Journal of Human Resource, 37*(4),752-777. https://doi.org/10.2307/3069616

Maharani, A.,& Putro, N. (2020). Item analysis of English final semester test. *IndonesianJournal of EFL and Linguistics*, *5*(2), 491-504.http://dx.doi.org/10.21462/ijefl.v5i2.302

Mahroof, A.,& Saeed, M. (2021). Evaluation of question papers by board of intermediate and secondary education using item analysis and Blooms taxonomy. *Bulletin of Education and Research December, 43*(3), 81-94. Retrieved from: https://eric.ed.gov/?q=item+analysis+&ft=on&id=EJ1341112

Muhson, A., Lestari, B., Supriyanto.,& Baroroh, K. (2017). The development of practical item analysis program for Indonesian teachers. *International Journal of*

Instruction, 10(2), 199-210. Retrieved from: http://www.e-iji.net/dosyalar/iji_2017_2_13.pdf

Malec, W.,&Krzeminska-Adamek, M. (2020). A practical comparison of selected methods of evaluating multiple-choice options through classical item analysis.*Practical Assessment, Research, and Evaluation*: *25*,Retrieved from: https://scholarworks.umass.edu/pare/vol25/iss1/7

Nana, E. (2018). *An Analysis of English Teacher-Made Tests. State University of Makasar.*

Pradanti, S., Martono, M., & Sarosa, T. (2018). An Item Analysis of English Summative Test for The First Semester of The Third Grade Junior High School Students in Surakarta. *English Education Journal*, *6*(3), 312-318.https://doi.org/10.20961/eed.v6i3.35891

Rudner, L.M.,&Schafer, W.D. (2002).*What Teachers Need to Know about Assessment*. Washington DC: National Education Association.

Salwa, A. (2012). *The Validity, Reliability, Level of Difficulty and Appropriateness of Curriculum of the English Test*. Diponegoro University.

Suharsimi, A. (2010). *Dasar-Dasar Evaluasi Pendidikan*, Jakarta: Bumi Aksara.

Santyasa, I. W. (2005). *Analisis Butir dan Konsistensi Internal Tes*. Retrieved from: http://johannes.lecture.ub.ac.id.

Setiyana R. (2016). Analysis of summative tests for English. *EEJ7*(4), 433-447.https:// 10.35308/ijelr.v2i2.2781

Semiun, T. T., & Luruk, F. D. (2020). The quality of an English summative test of a public junior high school, Kupang-NTT. *English Language Teaching Educational Journal, 3*(2), 133-141.https://doi.org/10.12928/eltej.v3i2.2311

Tarrant, M., Ware, J.,& Mohammed, A.M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med Educ 9*, 40. Retrieved from: https://doi.org/10.1186/1472-6920-9-40

Toksöz, S., & Ertunç, A. (2017). Item analysis of a multiple-choice exam. *Advances in Language and Literary Studies, 8*(6), 141-146. https://doi.org/10.7575/aiac.alls.v.8n.6p.140

Wells, C. S., & Wollack, J. A. (2003). *An Instructor's Guide to Understanding Test Reliability*. Madison: University of Wisconsin

Wisrance, M. W., & Semiun, T. T. (2020). LOTS and HOTS of teacher-made test in junior high school level in Kefamenanu*. Journal of English Education. 6* (2), 62–76. https://doi.org/10.30606/jee.v6i2.574

Wiyasa, P. I., Laksana, I K. D., Indrawati., & Mas, N. K. L. (2019), Evaluating quality of teacher-developed English test in vocational high school: content validity and item analysis. *Education Quarterly Reviews, 2*(2), 344-356. https://doi.org/10.31014/aior.1993.02.02.67