A Study on the Quality of Mersin University School of Foreign Languages English Preparatory Classes' Final Exam

Ulaş KAYAPINAR*

Abstract – The purpose of this study is to reveal the quality of the measurement instrument used as the final exam which was implemented at Mersin University School of Foreign Languages at the end of 2002-2003 academic year considering its psychometric characteristics such as item difficulty index, item discrimination index, and test reliability. The findings related to the item discrimination index indicated that 4 items (0,11-0,20) used in the exam were very difficult, 11 items (0,24-0,36) were difficult, 25 items (0,41-0,59) were moderately difficult, 33 items (0,62-0,80) were easy, and 21 items (0,82-0,95) are quite easy. The findings related to the item discrimination index indicated that 22 items could not discriminate the students who acquired the intended behaviours and who did not. Besides, 49 items could discriminate the examinees who had the necessary knowledge for responding to the items correctly. The reliability of the measurement instrument was found as 0,89. It can be stated that the measurement instrument is adequately reliable.
Key words: Measurement instrument, validity, reliability.

Özet – Mersin Üniversitesi Yabancı Diller Yüksekokulu İngilizce Hazırlık Sınıfları Final Sınavının Niteliği Üzerine Bir Çalışma – Bu çalışmanın amacı Mersin Üniversitesi Yabancı Diller Yüksekokulu 2002-2003 eğitim öğretim yılında İngilizce Hazırlık sınıfları final sınavı olarak uygulanan ölçme aracının niteliğini madde güçlük indeksi, madde ayırıcılık gücü indeksi ve test güvenirliği gibi psikometrik özellikleri bakımından ortaya koymaktır. Bulgular, sınavdaki 4 maddenin çok zor (0,11-0,20), 11 maddenin zor (0,24-0,36), 25 maddenin orta (0,41-0,59), 33 maddenin kolay (0,62-0,80), 21 adet maddenin de çok kolay (0,82-0,95) maddeler olduğunu ortaya koymuştur. Madde ayırıcılık gücüne ilişkin bulgular, 22 maddenin, o maddelerle ölçülmek istenen davranışlara sahip olan ve olmayan öğrencileri ayırt edemediğini göstermektedir. Sınavdaki 49 maddenin, o maddeleri cevaplamak için gerekli davranışa sahip olan ve olmayan öğrencileri ayırt edebildiği bulgulanmıştır. Ölçme aracının KR-20 güvenirlik katsayısı 0,89 olarak hesaplanmış olup sınavın yeterli güvenirlikte olduğu ifade edilebilir.
Anahtar kelimeler: Ölçme aracı, geçerlik, güvenirlik.

Introduction

Since education is a purposeful process and its purpose is to make the students learn predetermined necessary behaviours, a close control of both developments and

* Ulaş Kayapınar, research assistant, Mersin University Faculty of Education, Department of English Language Teaching. <ukayapinar@mersin.edu.tr>.

Mersin Üniversitesi Eğitim Fakültesi Dergisi, Cilt 2, Sayı 1, Haziran 2006, ss. 1-11.
Mersin University Journal of the Faculty of Education, Vol. 2, Issue 1, June 2006, pp. 1-11.

deficiencies is required in the teaching process. For this reason, the components of measurement and assessment in educational curriculums are used in different phases of educational process. In the eyes of educators who keep the importance of the process in mind because wrong decisions can lead drawbacks, the necessity of eligible measuring instruments which are used for making decisions can be seen clearly. Tests can progress the curriculums by shaping the intentions and expectations of both the students and the teachers, and they join all parts of a curriculum such as cohesion, purpose, and control (Brown, 1995).

Especially at universities in Turkey, curriculums of preparatory classes are developed for teaching English language intensively. In the curriculums of preparatory classes, various tests are developed and used extensively. These tests have powerful roles in many people's lives in the educational process and they are at significant transitional moments in education and beyond. Teachers rely on the information provided by the tests to make important decisions (McNamara, 2000). Rogers (1976) states that curriculum is a broad concept. It includes all the teaching and learning activities in which students take part with the support of the school. This means the description of the things to be learnt, the ways of learning these things, and the ways teachers give support in this process with the help of necessary materials, attributes, and methods of assessment (Johnson, 1989). At this point, the valid and reliable measuring instruments used in the English preparatory classes' curriculums have great importance in the educational process because they are conducive to the measurement and assessment, and the control of the educational process.

With this respect, the final examination which was implemented in the English preparatory classes of Mersin University School of Foreign Languages was examined according to its psychometric characteristics in order to reveal whether qualified individuals who are proficient in English are chosen by using the particular exam in a quality seeking global arena. Concerning this, in order to reveal the possible inadequacies in the particular exam and to present possible suggestions for the better, the characteristics indicating its quality such as having the properties of validity and reliability were tried to be examined by means of the research questions below.

How are the psychometric characteristics of the English preparatory classes' final exam implemented at Mersin University School of Foreign Languages at the end of 2002-2003 academic year?

a. How are the item difficulty index levels of the items in the exam?
b. How are the item discrimination index levels of the items in the exam?
c. How is the reliability of the exam?

Method

In this study, the psychometric characteristics of the final exam implemented in Mersin University School of Foreign Languages at the end of the 2002-2003 academic year

were tried to be revealed. There is not any data supplied from any sample which is generalized about a population; therefore, a present occasion was tried to be laid open. In this way, this research can be called as a fundamental research.

The study is not connected to any sample or population. Instead, a study group from which the data were supplied can be defined as one hundred and fifty seven students who had taken the final exam of Mersin University School of Foreign Languages English preparatory classes during the 2002-2003 academic year.

The 1st, 2nd, and 3rd sub-questions of the study are oriented towards the item difficulty, the item discrimination index levels of the items in the exam, and the reliability of the exam. In order to find answers to these questions, the item and test scores supplied from the responses of the students in Mersin University School of Foreign Languages 2002-2003 English preparatory classes' measurement instrument used as the final exam were analyzed.

As item statistics, item difficulty index (p_j) was computed for the dichotomous scores corresponding to two categories (1-0) (Baykul, 2000).

For item discrimination index, point-biserial correlation coefficients which give the item validity measures were computed. The point biserial correlation is used as the correlation between success and failure on the item and score (Thorndike, 1982), in other words, it shows the degree whether the item can discriminate the individuals who pass and fail according to one item (Baykul, 1999).

In order to measure the reliability of the test, *Kuder-Richardson* (KR-20) correlation coefficient was computed for binary scoring as one of the test statistics (Thorndike, 1982). Besides, the mean, median, mode, standard deviation, skewness, kurtosis, range, and minimum and maximum scores of the data gathered from the exam were computed as descriptive statistics of the test with the help of the SPSS 12.0 statistics programme.

Findings and Discussion

First of all, it is apt to give the test statistics of the exam with the descriptive statistics (Table 1) before the questions are given.

Table 1 indicates the number of the individuals (157) who were given the test (N). The mean of all scores (58,79) is also computed. Median, which is the score in the middle of the distribution, is seen as 59,00. Mode which is the score with the highest frequency is also 59,00. The standard deviation of the test which indicates the deviation from the mean of the distribution is 12,59. Skewness (-0,150) shows a slight skewed distribution to the left and it means that the items are below the average difficulty level. Besides, kurtosis (-0,605) states a low variability of the scores in the test. Range means the difference between the highest score and the lowest score (53,00). The minimum score is 30,00, and the maximum score is 83.00.

Table 1: *2002-2003 English Preparatory Classes' Final Exam Test Statistics*

N	157	**Skewness**	-0,150
Mean	58,79	**Kurtosis**	-0,605
Median	59,00	**Range**	53,00
Mode	59,00	**Minimum**	30,00

In Table 1, it is seen that mode and median have the same values (59,00), and mean (58,79) has a very close value to mode and median. With respect to these findings, it can be said that the scores of the final exam are likely to be seen in normal distribution. Kurtosis (-0,605) and skewness (-0,150) coefficients have negative values. These values give the information of a skewed distribution to the left and the distribution is slightly flat. Besides, the closeness of the distribution values to 0,00 supports that it almost resembles the normal distribution. Standard deviation (12,59) states that the variability of the distribution is narrow. With this respect, it can be stated that the distribution of the final exam scores appears quite homogenously. The results in connection to these findings can be observed from the distribution graphic in Figure 1.

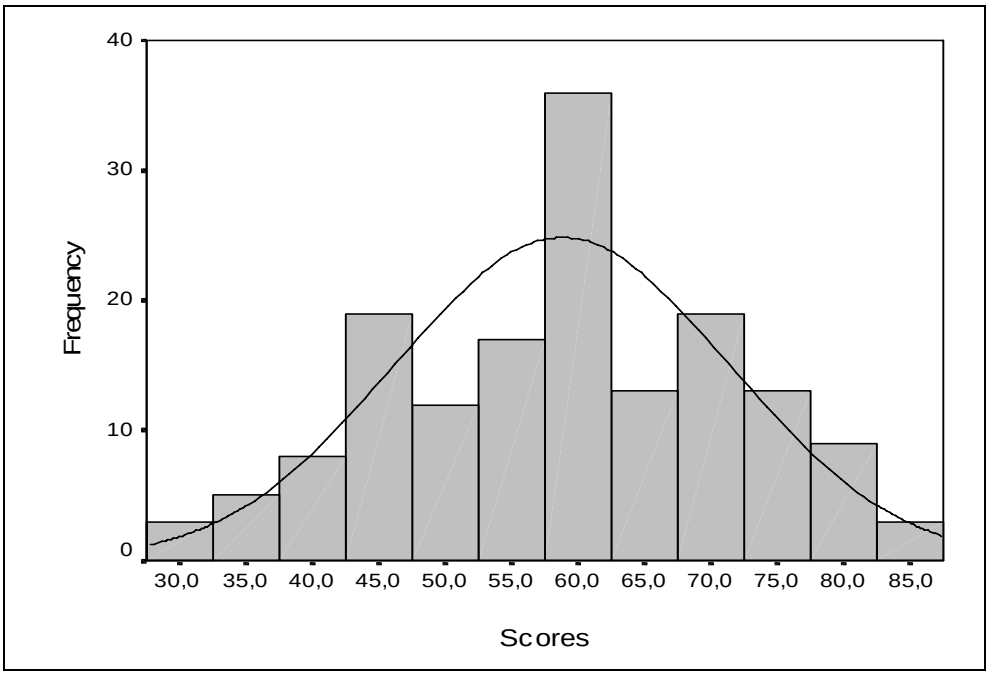

Figure 1: *The Distribution of Obtained Scores of 2002-2003 English Preparatory Classes' Final Exam*

As seen in Figure 1, the distribution of the final exam scores is similar to normal distribution. Beside that the mode value of the distribution is about sixty, the scores of forty five and seventy are seen to have more frequency than the other scores. However, the distribution has a single mode. As seen again in the graphic, the distribution is quite homogenous.

Findings and Discussion of Research Questions "a" and "b":

"How are the item difficulty index levels (p_j) of the items in the exam?"

"How are the item discrimination index levels (r_{jx}) of the items in the exam?"

To answer these questions, the findings of the item difficulty indexes are given in Table 2 in the first step.

In Table 2, item numbers are given in the first column. The second column includes the item difficulty index (p_j) which indicates the percentage of examinees who respond to the particular item correctly (Thorndike, 1982). When the item difficulty indexes are examined in Table 2, it can be observed that the items numbered 22, 31, 69, and 91 have values between 0.00 and 0.20. The findings indicate that these items are very difficult. The values of the items numbered 1, 7, 11, 12, 13, 41, 72, 73, 80, 85, and 87 differ between 0.20 and 0.40. These items can be called as difficult ones. The item difficulty index values of the items numbered 3, 4, 5, 6, 8, 14, 20, 24, 32, 34, 35, 40, 42, 44, 49, 51, 55, 56, 67, 74, 75, 82, 83, 84, and 94 are between 0,40 and 0,60. These items can be said to be moderately difficult ones. Moreover, the items numbered 2, 9, 10, 15, 16, 18, 19, 21, 23, 28, 29, 36, 37, 38, 43, 52, 53, 54, 57, 58, 59, 60, 68, 70, 71, 77, 81, 86, 88, 89, 90, 92, and 93 have values between 0.60 and 0.80; therefore, they can be called as easy items. Finally, the values of the items numbered 17, 25, 26, 27, 33, 39, 45, 46, 47, 48, 50, 61, 62, 63, 64, 65, 66, 76, 78, 79, and 95 differ between 0,80 and 0,97. In this way, these items are quite easy in connection to the item difficulty index.

Taken to the importance of better comments on item statistics and the decisions related to the items in a complete way, item difficulty and item discrimination indexes are examined together. With this respect, the findings of the item difficulty indexes are explained in the first step. After that, the items with high quality are tried to be revealed related to both item difficulty and item discrimination indexes. The item discrimination index values are given in Table 3.

In Table 3, item numbers are given in the first column. The second column points out the item discrimination index (r_{jx}) which is analyzed with the equation of the point biserial correlation coefficient.

Table 3 indicates that there are twenty-two items which have the r_{jx} value below 0,20 and there are twenty-four items which have the r_{jx} value between 0,20 and 0,30 ac-

Table 2: *2002-2003 English Preparatory Classes Final Exam Item Difficulty Indexes*

Item no	p_j	***Item no***	p_j	***Item no***	p_j	***Item no***	p_j	***Item no***	p_j	***Item no***	p_j
1	0.29	**17**	0.83	**33**	0.82	**49**	0.55	**65**	0.88	**81**	0.66
2	0.72	**18**	0.68	**34**	0.59	**50**	0.86	**66**	0.92	**82**	0.48
3	0.51	**19**	0.75	**35**	0.46	**51**	0.55	**67**	0.54	**83**	0.43
4	0.49	**20**	0.5	**36**	0.78	**52**	0.76	**68**	0.75	**84**	0.55
5	0.51	**21**	0.66	**37**	0.68	**53**	0.8	**69**	0.17	**85**	0.28
6	0.51	**22**	0.11	**38**	0.72	**54**	0.62	**70**	0.69	**86**	0.65
7	0.31	**23**	0.71	**39**	0.88	**55**	0.41	**71**	0.71	**87**	0.34
8	0.51	**24**	0.58	**40**	0.48	**56**	0.41	**72**	0.25	**88**	0.76
9	0.65	**25**	0.82	**41**	0.36	**57**	0.71	**73**	0.31	**89**	0.63
10	0.75	**26**	0.95	**42**	0.55	**58**	0.73	**74**	0.49	**90**	0.68
11	0.24	**27**	0.9	**43**	0.67	**59**	0.78	**75**	0.5	**91**	0.2
12	0.29	**28**	0.64	**44**	0.42	**60**	0.68	**76**	0.82	**92**	0.8
13	0.31	**29**	0.8	**45**	0.88	**61**	0.96	**77**	0.62	**93**	0.73
14	0.58	**30**	0.36	**46**	0.89	**62**	0.83	**78**	0.85	**94**	0.58
15	0.76	**31**	0.2	**47**	0.97	**63**	0.83	**79**	0.85	**95**	0.92
16	0.62	**32**	0.59	**48**	0.87	**64**	0.82	**80**	0.34		

cording to the item discrimination index which indicates item validity measures. The item numbers which have the r_{jx} value below 0,20 are 23, 24, 26, 30, 31, 33, 36, 39, 41, 43, 44, 45, 46, 51, 55, 56, 61, 66, 80, 87, 91, and 95. The findings indicate that these items cannot discriminate the students who acquired the intended behaviours and who do not acquire those behaviours related to the property asked to be measured. In another way, the items cannot discriminate the students who acquire the intended behaviours and who do not. Having negative values of the items numbered 51 (III. A. 31.) and 61 (III. B. 41; in the order of items, -0,08 and -0,02) related to the item discrimination index reveals that the items cannot measure the intended property. It can be stated that they try to measure some other properties instead of the intended ones. The check of the items which should be extracted from the test according to the item discrimination index levels was made with the help of 7 lecturers of English Language Teaching Department in Mersin University.

Some examples are given below:

Choose the correct answer.

51 (III. A. 31.). ____ of my parents seem happy about my leaving home to go abroad.

a. Both b. Neither c. Some d. None

Table 3: *2002-2003 English Preparatory Classes' Final Exam Item Discrimination Indexes*

Item no	r_{jx}	***Item no***	r_{jx}	***Item no***	r_{jx}	***Item no***	r_{jx}	***Item no***	r_{jx}	***Item no***	r_{jx}
1	0.29	**17**	0.32	**33**	0.19	**49**	0.49	**65**	0.34	**81**	0.31
2	0.38	**18**	0.39	**34**	0.22	**50**	0.35	**66**	0.18	**82**	0.29
3	0.52	**19**	0.38	**35**	0.44	**51**	-0.08	**67**	0.42	**83**	0.37
4	0.5	**20**	0.39	**36**	0.17	**52**	0.29	**68**	0.22	**84**	0.35
5	0.49	**21**	0.35	**37**	0.37	**53**	0.2	**69**	0.26	**85**	0.24
6	0.56	**22**	0.34	**38**	0.45	**54**	0.28	**70**	0.25	**86**	0.33
7	0.47	**23**	0.11	**39**	0.18	**55**	0.14	**71**	0.25	**87**	0.18
8	0.37	**24**	0.13	**40**	0.21	**56**	0.19	**72**	0.34	**88**	0.4
9	0.51	**25**	0.24	**41**	0.17	**57**	0.37	**73**	0.26	**89**	0.43
10	0.45	**26**	0.17	**42**	0.43	**58**	0.42	**74**	0.47	**90**	0.38
11	0.32	**27**	0.24	**43**	0.1	**59**	0.21	**75**	0.27	**91**	0.01
12	0.5	**28**	0.53	**44**	0.19	**60**	0.47	**76**	0.43	**92**	0.35
13	0.51	**29**	0.25	**45**	0.02	**61**	-0.02	**77**	0.3	**93**	0.27
14	0.35	**30**	0.01	**46**	0.12	**62**	0.36	**78**	0.38	**94**	0.21
15	0.39	**31**	0.07	**47**	0.22	**63**	0.24	**79**	0.42	**95**	0.02
16	0.26	**32**	0.26	**48**	0.41	**64**	0.4	**80**	0.05		

Although the correct choice of the item is “d” according to the answer key, it can be seen that the choice “a” can also be correct syntactically. Relatively, the item does not serve its construction aim and it naturally cannot measure a specific intended property.

Choose the correct answer.

61 (III. A. 41.).

A. What kind of films do you like best?

B. _____

a. I can’t stand horror films.
b. Why don’t we watch a video?
c. I love adventure movies.
d. I wouldn’t like to watch TV on a beautiful day like this.

The correct choice of this item is “c” according to the answer key but the experts commonly state this item can also have two probable answers according to the context of situation because one of the distractors, the choice “a”, can work in a more different way than the others as an available response in the spoken context. In this way, this item also does not work for its aim in a specific direction. Therefore, neither of the

items can discriminate the students who acquire the intended behaviours and who do not. This kind of items in the test could also decrease the reliability coefficient.

Moreover, the items which have the r_{jx} value between 0,20 and 0,30 are 1, 16, 25, 27, 29, 32, 34, 40, 47, 52, 53, 54, 59, 63, 68, 69, 70, 71, 73, 75, 82, 85, 93, and 94. The findings indicate that these items should be reexamined and redesigned. The validity problem of these items can be originated from the item stem**,** the choices (the correct answer and the distractors), answer key, etc. They should be examined in order to increase the item validity level of these items. Some of these items are given below:

Choose the correct answer.

40 (III. A. 20.). Sue ___ very strangely these days. Do you know what her problem is?

a. has been acted b. is acting c. will act d. acts

63 (III. B. 43.).

A. How old is that building do you think?
B. _____

a. There is a new cafe on the top floor.
b. It looks as if it is sort of ghostly.
c. I don't know. It looks modern.
d. It looks like a palace.

The items (40 and 63) seem to have more than one correct choice ("b" and "d" for 40; "b" and "c" for 63) according to the context of situation. When these items are also renewed or redesigned and the test is implemented again, the reliability coefficient of the test is supposed to change in a positive way.

Besides, the items which have the r_{jx} value higher than 0,30 are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 28, 35, 37, 38, 42, 48, 49, 50, 57, 58, 60, 62, 34, 65, 67, 72, 74, 76, 77, 78, 79, 81, 83, 84, 86, 88, 89, 90, and 92. These items are seen to measure the intended property successfully. In other words, they can discriminate the examinees who acquire the intended behaviour and who do not quite well. It can also be stated that these items are suitable for being used in the test without any reexamining or renewing.

The items which have values between 0,20 and 0,30 related to the item discrimination index should be reexamined and improved in order to join the test. The items which have values higher than 0,30 could be included in the test without any reexamining and improving. Moreover, normal distribution should be taken into consideration according to the item discrimination index. It is supposed to be normal or similar to normal (Tekin, 2000). In other words, the items should be chosen such as very difficult, difficult, average, easy, and very easy among the items which have higher values than 0,20 so as to resemble normal distribution. Examining Figure 2 could be helpful for a better understanding.

In Figure 2, item difficulty index levels are presented on the horizontal axis and item discrimination indexes are shown on the vertical axis. The numbers which are seen in the distribution related to the item difficulty (p_j) and item discrimination (r_{jx}) index levels refer to the item numbers. In the figure, the distribution of the items which have item discrimination values under 0,20, the items which have item discrimination values between 0,20 and 0,30, and the items which have item discrimination values higher than 0,30 could be seen connected to the item difficulty index levels.

Item difficulty indexes of the relevant items can be observed between 0.20 and 0.90 in the distribution graphic. Item discrimination index values of the items which are higher than 0,30 change approximately between 0,10 and 0,90 in the graphic related to the item difficulty index. By examining the graphic in the Figure 2, the distribution of the items which have item discrimination values higher than 0,20 can be observed and the suitable items which have the distribution similar to the normal distribution according to the item difficulty indexes could be chosen.

Findings and Discussion of Research Question "c":

"How is the reliability of the exam?"

The reliability of the exam was computed with *Kuder-Richardson* reliability coefficient (KR-20). The KR 20 reliability coefficient of the test was found as 0,89. In fact, when the twenty two items which have reliability values under 0,20 were removed from the test and the KR-20 reliability coefficient of the test was computed with the rest of the items, the value of the test with 73 items was found as 0,91. This KR-20 coefficient seems higher than the one (0,89) which is supplied from the test including 95 items. The number of the items can affect the reliability. Moreover, the reliability coefficient increases as the number of the items increase in certain limits (Anastasi, 1982). If new items can be included instead of the items removed because of the lower reliability coefficients, a higher reliability coefficient than 0,91 is supposed to be obtained as a result. However, it is obvious that a reliable tool can only be used if it is also valid (Erkuş, 2003). A high reliability coefficient is not sufficient alone without acceptable validity levels.

Conclusion

In this study, the psychometric characteristics of the English preparatory classes' final exam implemented at Mersin University School of Foreign Languages at the end of 2002-2003 academic year were tried to be examined. There are 49 items which have the r_{jx} value higher than 0,30 out of 95. These items are seen to measure the intended property in a successful manner. In another way, they can discriminate the examinees

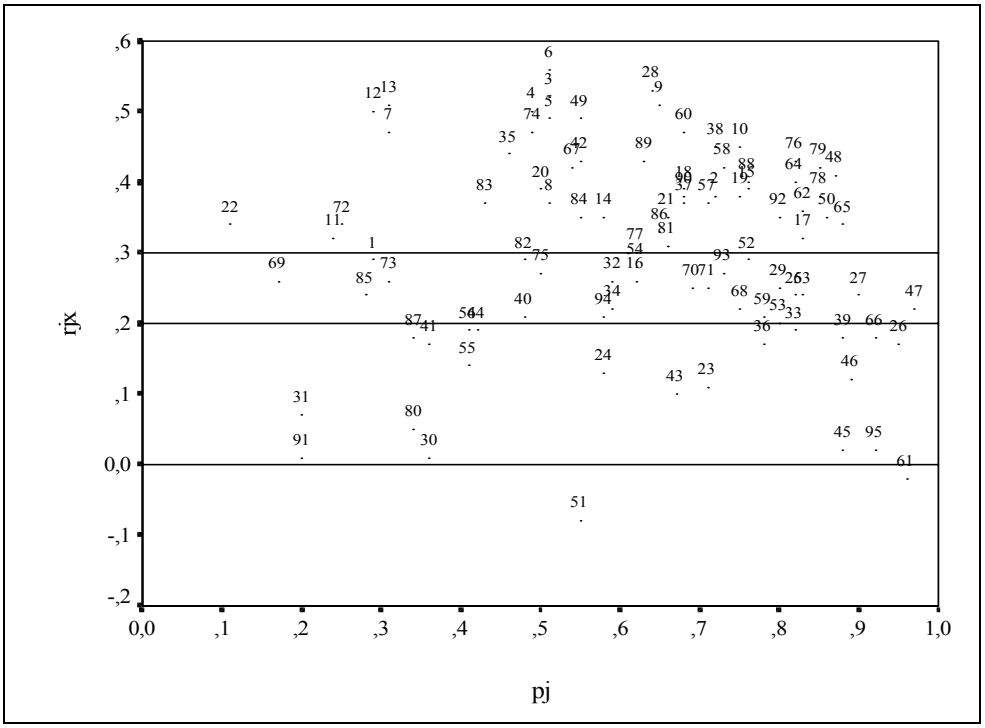

Figure 2: *2002-2003 English Preparatory Classes' Final Exam Measurement Instrument Item Difficulty and Item Discrimination Indices*

who know the answer and who do not know the answer. These items can be called as suitable to be used in the test without any reexamining or renewing. As another psychometric characteristic, the reliability of the exam was computed with *Kuder-Richardson* reliability coefficient (KR-20). The KR-20 reliability value of the test is 0,89 which can be called as a high reliability coefficient. The results show that 24 items in the exam should be reexamined and 22 items should be extracted from the test..

Concerning this, test development and curriculum evaluation teams can be established not only at the particular university which the study was held, but also at all educational institutions which play roles in making important decisions for people's educational lives and beyond. Those teams can include the participation of the teachers who individually teach English preparatory classes and some experts of measurement and evaluation fields. By this way, an item pool can be prepared which is full of valid and reliable items to be used in the tests without any doubt or incongruity; therefore, the individuals who are to be chosen as proficient in English can be better determined and the exams could prove their aims in a regular way.

Yet, taken the importance of the intensive study, labor, the amount of tests in English teaching classes programmes, significant transitional moments of people's lives in education and beyond, respect, and fairness, some studies have been held in different educational institutions (Kurtuluş, 2002; Tarman, 2002; Serpil, 2000; Osken, 1999; Osken, 1999; Ataman, 1999) in order to construct qualified tests for choosing qualified individuals. Moreover, similar studies could be standardized by all educational institutions because such studies help both educators and students to know that they achieve the necessities of a cornerstone in their lives without any questioning or excuse.

References

Anastasi, A. (1982). *Psychological testing*. USA: Macmillan Publishing.

Ataman, F. (1999). *Reliability and validity studies of METU School of Foreign Languages Department of Basic English June 1998 and September 1998 proficiency tests,* Unpublished MA Dissertation, METU, Ankara.

Baykul, Y. (2000). *Eğitim ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: ÖSYM Yayınları.

Baykul, Y. (1999). *İstatistik: Metodlar ve uygulamalar*. Ankara: Anı Yayıncılık.

Brown, J.D. (1995). *The elements of language curriculum*. Massachusettes: Heinle & Heinle.

Erkuş, A. (2003). *Psikometri üzerine yazılar*. Ankara: Türk Psikologlar Derneği Yayınları.

Johnson, R.K. (1989). *The second language curriculum*. Great Britain: CUP.

Kurtuluş, E. (2002). *The validity and reliability study of a pilot proficiency exam in 8th grades in METU development foundation school*, Unpublished MA Dissertation, METU, Ankara.

McNamara, T. (2000). *Language testing*. Oxford: OUP.

Osken, H. (1999). *An assessment of the validity of the midterm and the end of the course assessment tests administered at Hacettepe University, Department of Basic English.* Unpublished MA Dissertation, Bilkent University, Ankara.

Osken, H. (1999). *An investigation of the content validity and backwash effect of the end-of-term oral assessment test administered at Hacettepe University, Department of Basic English,* Unpublished MA Dissertation, Bilkent University, Ankara.

Serpil, H. (2000). *An assessment of the content validity of the midterm achievement tests administered at Anadolu University Foreign Languages Department,* Unpublished MA Dissertation, Bilkent University, Ankara.

Tarman, S. (2002). *Reliability and validity assessment and evaluation of Gazi Uni. department of music education "entrance-musical aptitude exams"*, Unpublished PhD Dissertation, Gazi Üniversity, Ankara.

Tekin, H. (2000). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı Yayınevi.

Thorndike, R.L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin Company.