Centre for the Greek Language

Statistical Analysis for the Certificate of Attainment in Greek 2009 Administration

Final Project Report

Spiros Papageorgiou
University of Michigan

Thessaloniki, Greece 2009

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

This report presents a statistical analysis of the Reading and Listening components of the Certificate of Attainment in Greek (Levels A, B, C and D), offered by the Centre for the Greek Language[1].

Examinee responses to reading and listening test items were collected at the University of Thessaloniki, Greece, from the 2009 administration of the test. Correct responses received one point, incomplete or incorrect responses received 0. The data were then analyzed with Classical Test Theory (Verhelst, 2004), using the computer programs Microsoft Excel and SPSS. The Classical Test Theory (CTT) statistics are explained in the report, however the interested reader can also consult widely-used language testing course books (inter alia, Alderson, Clapham, & Wall, 1995; Bachman, 2004). Classical Test Theory was preferred to Item Response Theory because it is accessible to a wider audience. However, it should be stressed that statistical information obtained through CTT is usually confined to the population that took the test. Therefore, the difficulty of an item might not be the same with a different test-taking population.

The analysis is presented in two main sections: Section 2 for reading and Section 3 for listening. Both sections are similarly structured, presenting the distribution of scores firstly and the item analysis secondly. It should be noted that raw (i.e. non-transformed) scores, are used. Therefore, if scores are reported to examine in a different format than raw scores (e.g. percentages or scaled scores), score distributions in the report might not be identical to the distribution of reported scores.

[1] http://www.greeklanguage.gr/

2 Reading component

2.1 Reliability and distribution of raw scores

The reading component of the Certificate of Attainment in Greek consists of 21 to 33 items, depending on the level (Table 2.1, Columns 1 and 2). Examinee numbers as can be seen in the third column (N) varied to 320 to 342. The remaining columns in the table show the following:

- Column 4: Cronbach's Alpha, an index that provides information about the internal consistency of the test. It is often used as an indicator of reliability, that is, how reliably a test separates students into different ability levels. Homogeneity of the test taking population and length of the test (if item quality remains constant) might affect Alpha. In general, the more heterogeneous the population and the longer the test (with items of the same quality), the higher the Alpha. Alpha indices of .85 and above are considered acceptable for high-stakes tests.
- Column 5: The arithmetic mean (average) of the scores obtained.
- Column 6: The Standard Deviation (SD), a statistic that indicates how much, on average, scores vary or deviate from the mean. In a normal distribution, approximately two thirds of the scores will fall within one SD.
- Columns 7 and 8: The minimum and the maximum scores are listed.
- Columns 9 and 10: Skewness and Kurtosis are indicators of the shape of distribution. If values of zero are observed, then the distribution is normal. In general, values between -2 and +2 indicate a reasonably normal distribution. Distributions are also graphically presented in histograms (Figure 2.1 to Figure 2.4). Positively skewed distributions indicate that more scores appear on the left-hand side of the histogram, i.e. more lower scores are observed. Negatively skewed distributions indicate that more scores appear on the right-hand side of the histogram, i.e. examinees primarily obtained high scores. Kurtosis indicates the degree of peakedness. Higher positive values signal a highly peaked (leptokurtic) distribution, whereas negative values signal a flat (platykurtic) distribution.

The results in this section suggest that the reliability of the reading component is acceptable, because the Alpha for all tests, apart from Level B, is between 87 and 90. For Level B, the lowest Alpha is observed most probably because the population was more homogeneous compared to those of the other levels. This homogeneity is indicated by the SD, with the one for Level B being the lowest, as well as by the low frequency of scores below 15 (Figure 2.3). The skewness and kurtosis figures suggest reasonably normal distribution for Levels C and D. As also illustrated in Figure 2.1 and Figure 2.2, the majority of scores For Levels A and B appears in the right-had side of the histograms. In other words, most examinees obtained very high scores.

Table 2.1 Reliability and distribution of scores for the reading component

Level	K	N	Alpha	Mean	SD	Min	Max	Skewness	Kurtosis
A	25	342	.92	21.92	4.76	0	25	-1.95	3.36
B	25	336	.80	22.50	2.97	9	25	-1.81	3.93
C	21	329	.88	15.86	4.74	3	21	-.61	-.75
D	33	320	.87	25.24	5.73	4	33	-.76	.26

Figure 2.1 Distribution of scores for Level A Reading

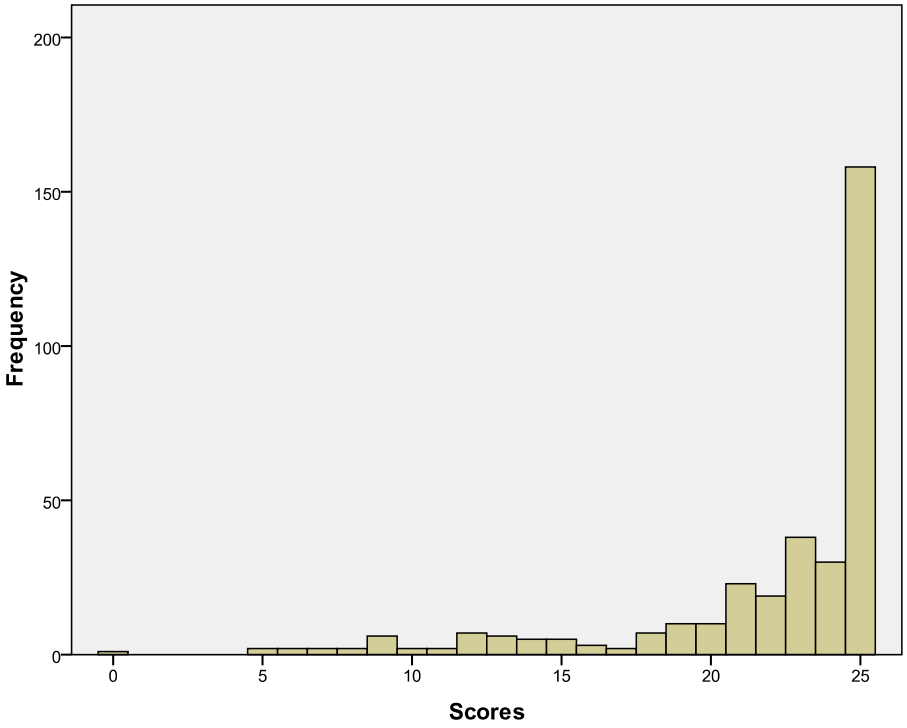


Figure 2.2 Distribution of scores for Level B Reading

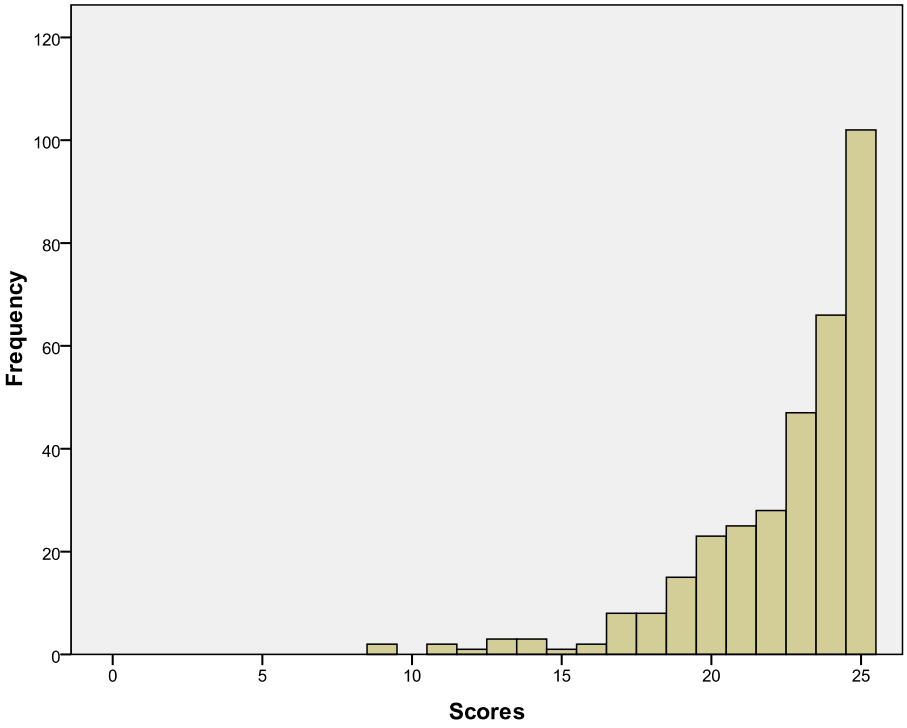


Figure 2.3 Distribution of scores for Level C Reading

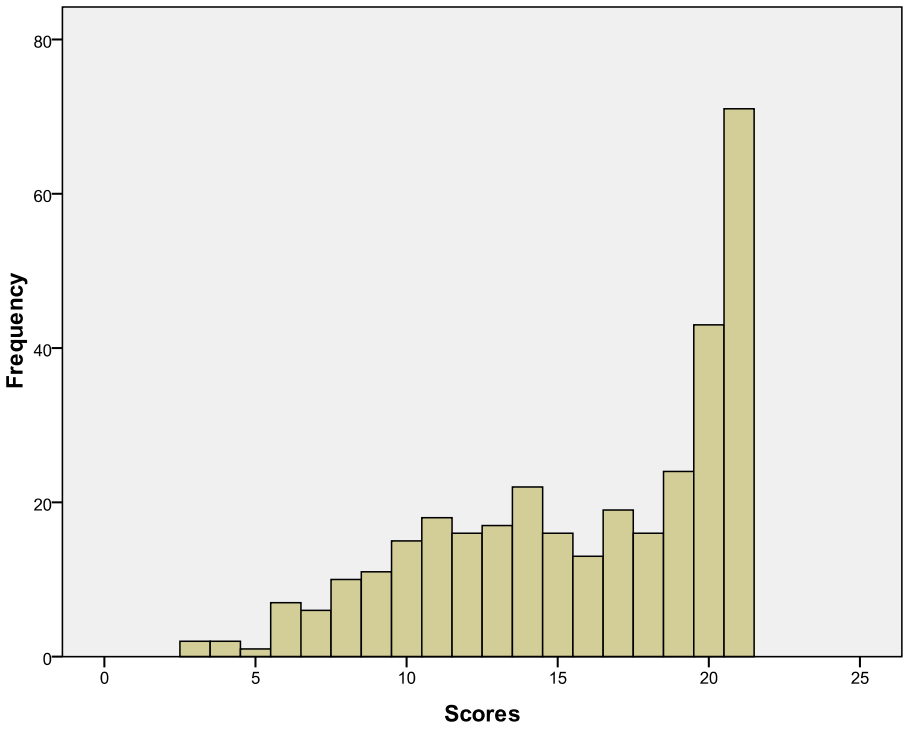
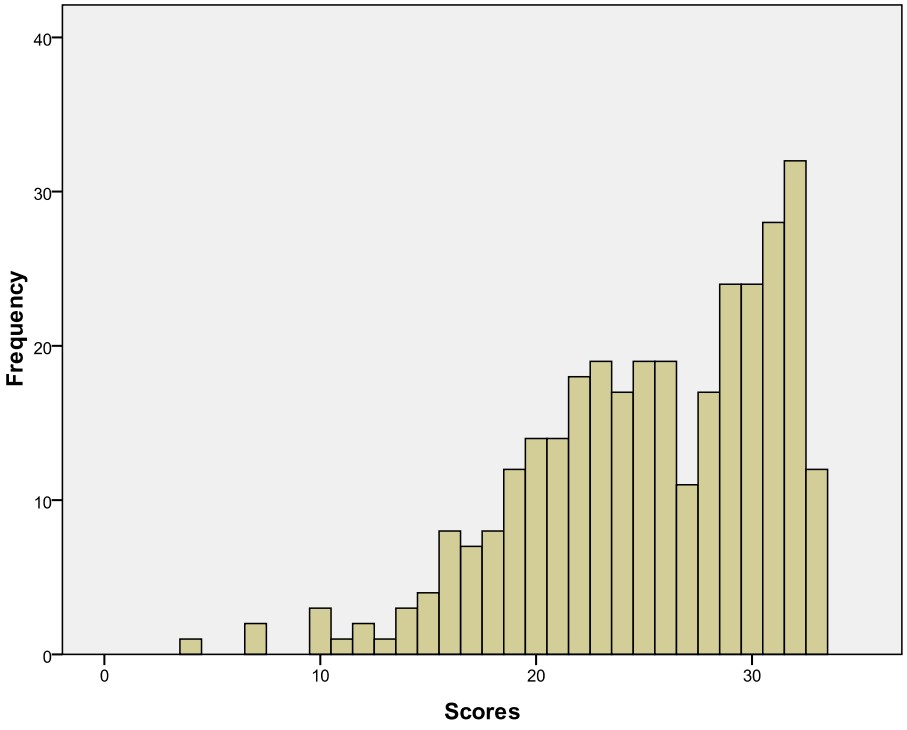


Figure 2.4 Distribution of scores for Level D Reading



2.2 Item analysis

Table 2.2 lists the item difficulty (in terms of percentage of people who answered items correctly) and item discrimination (in terms of corrected item-total correlation calculated by SPSS) for the four levels. The higher the difficulty value, the easier the item, as more examinees responded correctly. Discrimination values of .25 and above are indicators of a good quality item that can reliably separate test takers into higher and lower levels of ability. For multiple choice items, low discrimination might be attributed to a second possible answer or guessing. For open-ended items involving raters and a list of acceptable answers, low discrimination might be attributed to lack of internal consistency of the raters or non-inclusion of possible answers in the list of acceptable answers. Values below .25 in Table 2.2 are highlighted in bold for ease of reading.

The internal consistency of the test and the quality of the items are related. For example, the test with the lowest Alpha in Table 2.1 (Level B) also has the highest number of items with low discrimination. The tests with the highest Alpha (Levels A and C) have no items with discrimination values below .25. As mentioned earlier, the longer a test, the higher the reliability (if the quality of items remains the same). Because the Level D test is longer, the three items with low discrimination do not seem to have a notable impact on the Alpha value, which remains at .87.

As illustrated by the histograms in the previous section, Levels A and B were easy for the population. For level A, all but four items had a difficulty value of .82 and above, thus more than 80% of the examinees responded to these items correctly. Similarly, all but four Level B items had difficulty values of .83 and above.

Table 2.2 Item difficulty and discrimination for the reading component

	Level A		Level B		Level C		Level D	
	Difficulty	Discrim.	Difficulty	Discrim.	Difficulty	Discrim.	Difficulty	Discrim.
Item 1	.96	.38	.75	**.17**	.94	.38	.89	.29
Item 2	.93	.40	.93	**.16**	.85	.29	.90	.44
Item 3	.94	.50	.75	.33	.89	.38	.82	.33
Item 4	.96	.30	.97	**.22**	.76	.47	.73	.39
Item 5	.92	.42	.99	**.08**	.71	.41	.92	.31
Item 6	.92	.43	.93	.40	.91	.35	.93	.27
Item 7	.94	.32	.98	.34	.80	.42	.85	.42
Item 8	.94	.50	.99	**.18**	.83	.37	.83	.47
Item 9	.96	.36	.94	.25	.91	.29	.94	**.19**
Item 10	.95	.51	.90	.30	.80	.38	.98	**.14**
Item 11	.88	.51	.96	**.24**	.95	.28	.96	.37
Item 12	.93	.52	.98	**.19**	.55	.26	.92	.32
Item 13	.83	.69	.83	.47	.69	.43	.54	.26
Item 14	.78	.63	.96	.35	.59	.64	.74	**.09**
Item 15	.87	.59	.93	.35	.81	.51	.90	.34
Item 16	.84	.68	.85	.49	.55	.71	.69	.43
Item 17	.72	.57	.95	.36	.69	.72	.73	.35
Item 18	.88	.63	.87	.39	.60	.73	.87	.37
Item 19	.88	.67	.86	.53	.69	.53	.87	.32
Item 20	.77	.63	.76	.32	.68	.73	.76	.38
Item 21	.87	.66	.79	.43	.64	.72	.47	**.18**
Item 22	.77	.62	.91	.30			.75	**.24**
Item 23	.82	.65	.97	.41			.94	.28
Item 24	.85	.62	.83	.54			.68	.34
Item 25	.86	.66	.91	.39			.66	.40
Item 26							.60	.57
Item 27							.62	.53
Item 28							.54	.55
Item 29							.84	.46
Item 30							.61	.56
Item 31							.61	.61
Item 32							.61	.58
Item 33							.55	.61

3 Listening component

3.1 Reliability and distribution of raw scores

The listening component of the Certificate of Attainment in Greek consists of 25 to 33 items, depending on the level (Table 3.1, Columns 1 and 2). Examinee numbers as can be seen in the third column (N) varied from 320 to 358. The remaining columns in the table provide information similar to that in Section 2.1, that is, Cronbach's Alpha, mean, SD, minimum and maximum scores, skewness and kurtosis.

The results in this section suggest that the reliability of the listening component is acceptable for Levels A and D but lower for Levels B and C. The lower SD for these two levels and the lack of scores below 15 indicate a more homogeneous population, which explains the lower Alpha. The histograms (Figure 3.1 to Figure 3.4) in this section and the skewness and kurtosis figures also indicate a reasonably normal distribution only for Level D. The vast majority of the scores for the remaining levels are found in the right-hand side of the histograms; in other words very high scores were obtained.

Table 3.1 Reliability and distribution of scores for the listening component

Level	K	N	Alpha	Mean	SD	Min	Max	Skewness	Kurtosis
A	25	342	.91	20.82	3.55	0	25	-1.82	5.65
B	25	336	.74	22.05	2.50	7	25	-1.97	5.60
C	25	358	.65	22.50	2.29	0	25	-3.40	26.11
D	33	320	.82	25.03	5.05	6	33	-.719	.301

Figure 3.1 Distribution of scores for Level A Listening

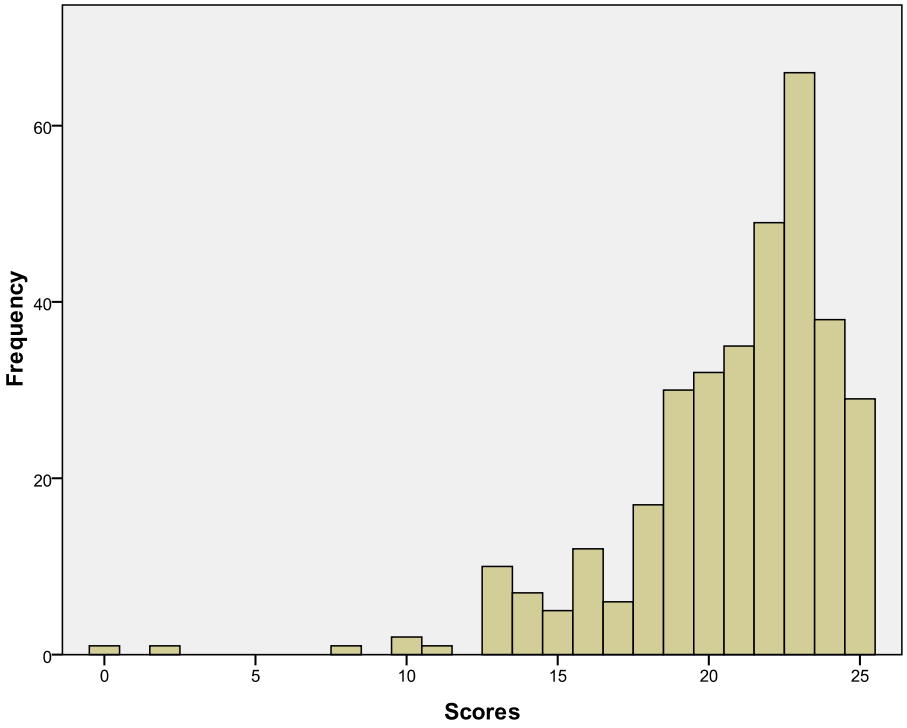


Figure 3.2 Distribution of scores for Level B Listening

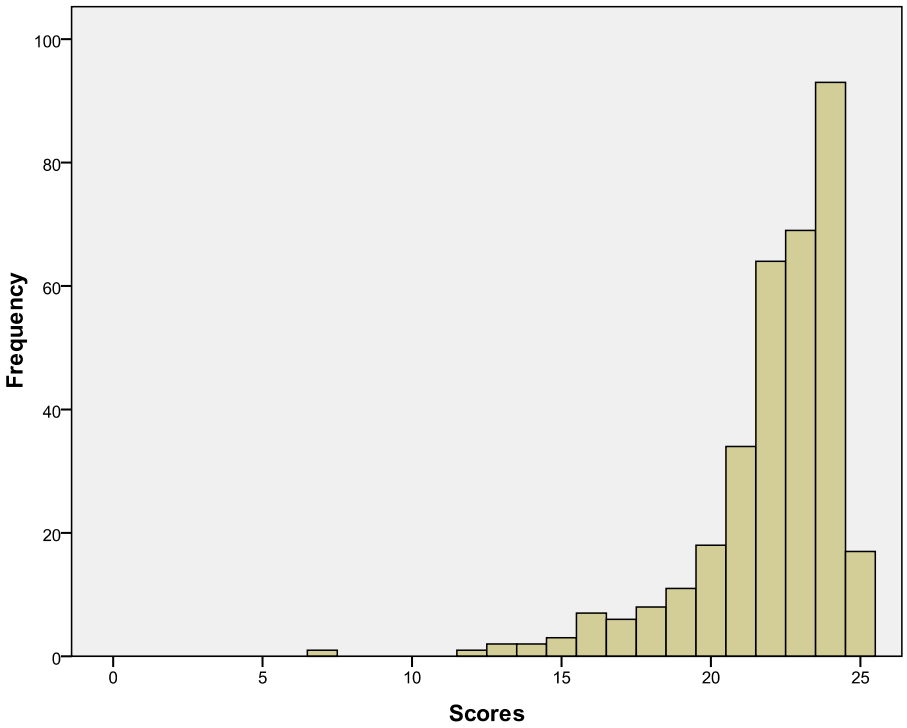


Figure 3.3 Distribution of scores for Level C Listening

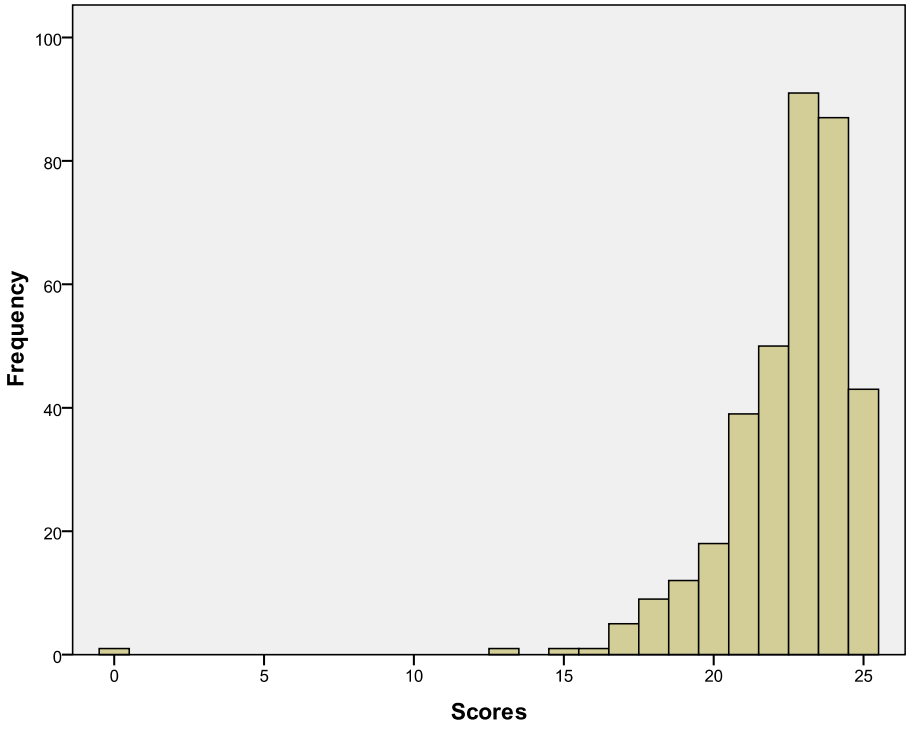
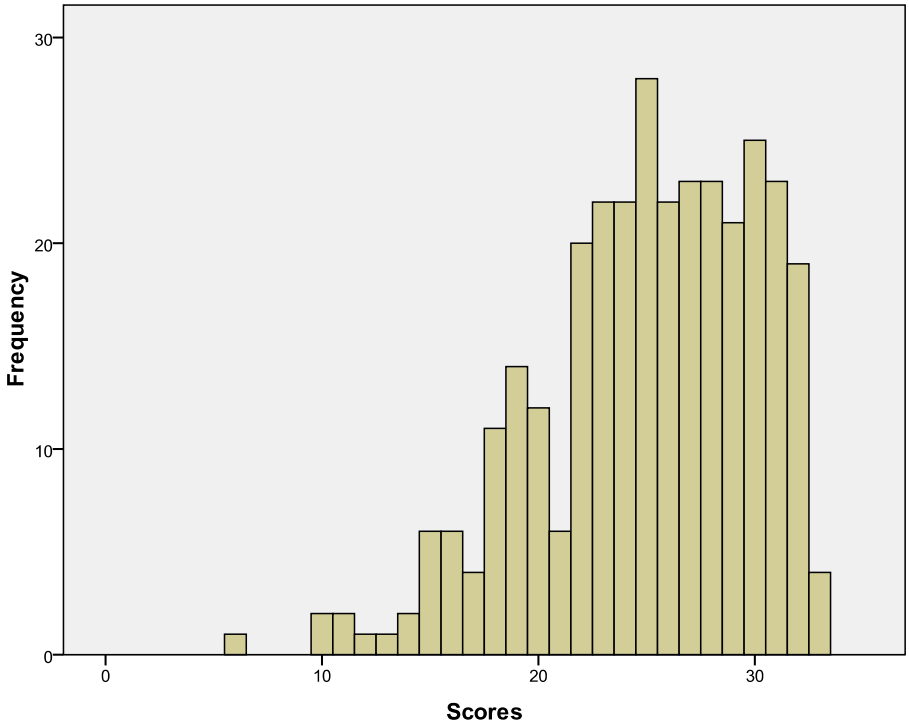


Figure 3.4 Distribution of scores for Level D Listening



3.2 Item analysis

Similarly to the statistics in Section 2.2, Table 3.2 lists the item difficulty (in terms of percentage of people who answered items correctly) and item discrimination (in terms of corrected item-total correlation calculated by SPSS) for the four levels. Discrimination values below .25 in Table 2.2 are highlighted in bold for ease of reading, as they indicate potentially problematic items.

Because the internal consistency of the test and the quality of the items are related, the test with the highest Alpha in Table 3.1 (Level A) also has the lowest number of items with low discrimination. Level C, with the lowest Alpha, has 10 out of 25 items with low discrimination. Because the Level D test is longer, the 12 items with low discrimination do not seem to have a notable impact on the Alpha value, which remains at .82.

As illustrated by the histograms in the previous section, Levels B and C were easy for the population. For level B, all but four items had a difficulty value of .82 and above, thus more than 80% of the examinees responded to these items correctly. Similarly, all but three Level C items had difficulty values of .80 and above.

Table 3.2 Item difficulty and discrimination for the listening component

	Level A		Level B		Level C		Level D	
	Difficulty	Discrim.	Difficulty	Discrim.	Difficulty	Discrim.	Difficulty	Discrim.
Item 1	.78	.47	.96	.37	.74	.26	.77	.25
Item 2	.78	.49	.96	.25	.96	**.16**	.97	**.19**
Item 3	.68	**.20**	.87	.26	.95	**.22**	.68	.37
Item 4	.75	**.22**	.94	.25	.97	.35	.83	.38
Item 5	.85	.50	.87	.33	.96	**.21**	.77	**.19**
Item 6	.68	.33	.96	**.04**	.95	.33	.95	**.17**
Item 7	.80	.48	.87	.26	.98	.28	.56	**.09**
Item 8	.82	.57	.94	.40	.95	.40	.92	**.23**
Item 9	.72	.37	.92	**.19**	.97	.25	.60	.27
Item 10	.91	.36	.82	**.21**	.97	.39	.97	**.13**
Item 11	.71	.41	.91	.31	.98	.39	.99	**-.01**
Item 12	.45	**.23**	.25	**-.08**	.96	.35	.34	**.07**
Item 13	.74	.35	.89	**.14**	.96	.28	.96	**.17**
Item 14	.96	.32	.66	**.14**	.95	.29	.58	.31
Item 15	.62	.26	.90	.34	.61	**.12**	.85	**.17**
Item 16	.95	.25	.92	.50	.85	.38	.92	**.20**
Item 17	.99	.41	.98	**.24**	.93	**.05**	.95	**.15**
Item 18	.98	.25	.95	.52	.80	**.08**	.81	.32
Item 19	.95	.41	.94	.29	.95	.40	.75	.34
Item 20	.97	.27	.97	.39	.59	**.10**	.73	.41
Item 21	.87	.31	.94	.47	.92	**.16**	.82	.43
Item 22	.96	.34	.84	**.17**	.90	**.10**	.83	.34
Item 23	.98	.34	.95	.32	.86	**.11**	.56	.48
Item 24	.99	.31	.98	**.13**	.93	.27	.47	.42
Item 25	.99	.33	.86	.26	.89	.32	.62	.53
Item 26							.51	.49
Item 27							.58	.35
Item 28							.74	.41
Item 29							.81	.45
Item 30							.76	.49
Item 31							.79	.43
Item 32							.79	.47
Item 33							.85	.38

4 Conclusion

This report presented a statistical analysis of the Reading and Listening components of the Certificate of Attainment in Greek (Levels A, B, C and D), offered by the Centre for the Greek Language.

Reliability was satisfactory for the reading component, with Level B having the lowest one. Items of Levels A and B also appeared to be relatively easy; therefore adding some more difficult items might something that the test developers want to consider. The addition of some more difficult items will probably result in a larger variety of scores and increase in the reliability for these two levels. Finally, the examination provider might want to explore whether Level B examinees should actually take Level C.

The reliability of the listening component was lower than the reliability of the reading component, particularly for Levels B and C, for which the existence of high scores suggested a homogeneous population and low item difficulty. As with the reading component, adding some more difficult items might result in higher reliability. Moreover, item analysis showed that items of low discrimination were more frequent in the listening component than the reading component. Extensive trialing of items might contribute to better discrimination.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Verhelst, N. (2004). *Classical Test Theory. Section C of the Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment.* Strasbourg: Council of Europe.