ANALYSIS OF MULTILEVEL ENGLISH LANGUAGE PROFICIENCY TESTS

A.A. Abbosov

Agency for Assessment of Knowledge and Competences under the Ministry of Higher Education,
Science and Innovations of the Republic of Uzbekistan,
100084, Tashkent., Bogishamol str., 12

Abstract. A multilevel English proficiency test is a test that assesses a test taker's English language skills at different levels of proficiency. This type of test is becoming increasingly popular in educational and workplace settings, as it allows for a more nuanced understanding of a test taker's language skills. This article provides an overview of the benefits and challenges of a multilevel English proficiency test and discusses the results of the multilevel English proficiency tests conducted by the Agency for Assessment of Knowledge and Competences. The test results are analyzed across different sections based on Classical test theory and IRT.

Keywords: Multilevel proficiency test, IRT, Rasch model, difficulty, ability, standard score, reliability.

Introduction

English proficiency tests are widely used to assess the language skills of non-native speakers of English. These tests are us ed in educational and workplace settings to determine language proficiency levels and to place test takers in appropriate language courses or job positions. A multilevel English proficiency test is a test that assesses a test taker's English language skills at different levels of proficiency. Multilevel testing is gaining more popularity in both academic and professional environments since it provides a more comprehensive and detailed insight into the language proficiency of the test taker.

A multilevel system of foreign language proficiency assessment was developed and put into practice by Assessment Agency (formerly known as State Testing Center) in 2022, based on the decree of the Cabinet of Ministers of the Republic of Uzbekistan No. 73 of February 16, 2022. During the months of March-December 2022, 11 exam sessions were organized and more than 45 thousand test takers participated in these tests.

Test results were analyzed using Classical test theory and IRT. As the results for listening and reading sections, latent ability scores based on the Rash model are reported in the form of standard scores. As for speaking and writing, the responses are evaluated by human raters against a pre-established scoring criteria. Cut-off scores for different levels are given in *Table 1*.

Level	Score
C1	65-75
B2	51-64
B1	38-50
Below B1	1-37

Table 1. Distribution of cut-off scores

1. Benefits and challenges of Multilevel tests

A multilevel proficiency test has several benefits over a traditional English proficiency test. One benefit is a comprehensive assessment of language skills. A traditional English proficiency test typically assesses language skills at a single level of proficiency, such as beginner, intermediate, or advanced. In contrast, a multilevel English proficiency test assesses language skills at multiple levels of proficiency, providing a detailed picture of a test taker's language skills. This leads to more accurate evaluations of language proficiency, which can help teachers and administrators make more informed decisions about placement and instruction [1].

Multilevel testing can lead to improved instruction and curriculum development. By providing a nuanced understanding of language proficiency, multilevel testing can help teachers and administrators identify areas of strength and weakness in language instruction. This can help inform curriculum development and instructional strategies, leading to better outcomes for students [2].

However, multilevel testing also presents some challenges in terms of test design and interpretation.

One of the main challenges of multilevel testing is test design. Tests must be designed to accurately measure proficiency at each level, while also allowing for comparisons across levels. This can be difficult to achieve, as different levels may require different types of questions or tasks. Additionally, tests must be designed to be fair and unbiased, regardless of the test-taker's level of proficiency [2].

Finally, multilevel testing presents challenges in terms of interpretation. Test results must be interpreted in a way that accurately reflects the test-taker's level of proficiency, while also allowing for comparisons across levels. This can be difficult to achieve, as different levels may have different scoring criteria or cut-off scores [1].

2. Equating process in the Rasch model

In order to ensure different versions of the test yield similar results and be comparable, a common scale must be created. There are different methods of creating a common scale for different forms of tests in Classical and Modern test theories. One of them is equating, which is the process of linking test scores from different forms of a test or different testing occasions to create a common scale for score interpretation [3]. Equating is a crucial process for educational and psychological assessments, as it allows for the comparison of test scores across different groups of test takers or different testing conditions. The Rasch model, a widely used item response theory model, provides a robust framework for equating test scores [4].

Equating in the Rasch model involves three main steps: (1) calibrating the test forms or testing occasions, (2) linking the test forms or testing occasions to create a common scale, and (3) evaluating the equating results.

Calibration involves estimating the item parameters (i.e., difficulty and discrimination) and the person parameters (i.e., ability) separately for each test form or testing occasion [5]. This is typically done using maximum likelihood estimation or Bayesian estimation methods.

Linking involves establishing the relationship between the test forms or testing occasions by aligning the item and person parameters on a common scale [6]. In this case, a test paper will consist of non-repeating (unique) and overlapping test items (*Table 2*). The overlapping items helps to create a common scale to ensure the parallelism of test forms when calculating test results [7].

Items\Versions	V1	V2	V3	V4
Overlapping items	6	6		
		6	6	
			6	6
	6			6
Unique items	23	23	23	23
Total	35	35	35	35

Table 2. Linking design sample for four test versions

Evaluation involves assessing the quality of the equating results. This can be done using various statistical methods, including the equating error, the standard error of equating, and the equating stability coefficient.

3. Analysis of test performance

The reliability in the assessment of writing and speaking skills requiring human participation is constantly monitored using Routine double check method [8]. Pearson's correlation coefficient is used in reliability analysis.

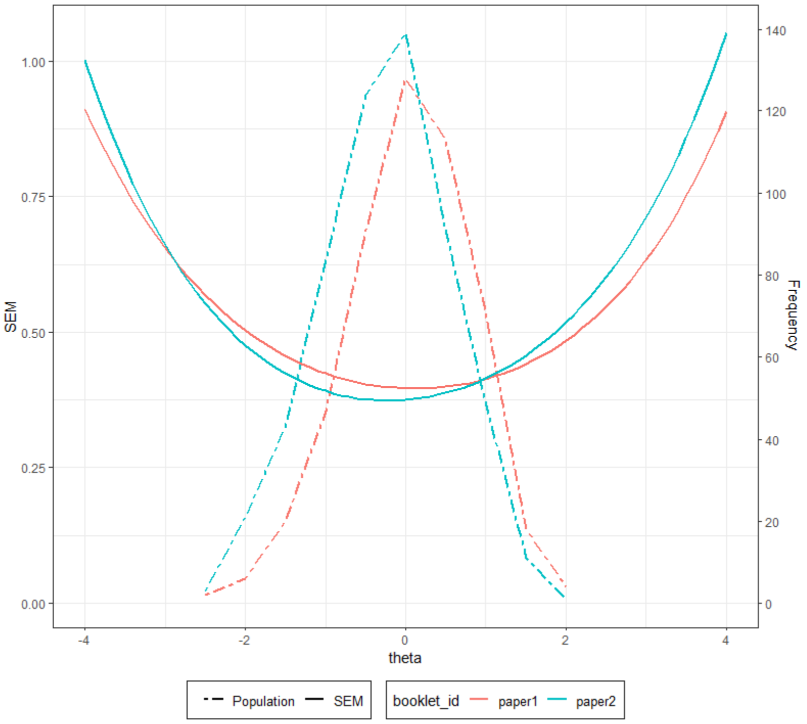
Mean correlations for Writing and Speaking are as follows:

Writing	0.843
Speaking	0.797

A Pearson correlation coefficient greater than 0.7 has been reported in the literature to indicate high inter-rater reliability [9].

Reliability in Listening and Reading is ensured by taking certain measures, including following the test specifications and the linking design requirements, as well as monitoring the quality of items using Classical test theory and IRT.

In order to improve the multi-level test system, cooperation with CITO experts is underway. The validity and reliability of the test results are being studied based on the Rasch model [10]. Together with the CITO expert, the results of the test conducted in March were analyzed (*Graph 1*) and it was noted that the reliability of the test results is high.



Graph 1. Relationship between cut-scores and test SEM

In this graph, the parabola lines represent the standard error of the two versions of the test, and the dashed lines represent the recorded latent ability scores. As can be seen from the graph, most of the scores are reported in the interval where the error value is small.

Moreover, internal consistency reliability is measured for each version of the test. Internal consistency reliability is a measure of the consistency with which an assessment tool or test measures a construct. It is a statistical measure that assesses how consistently the items or questions within a test measure the same underlying construct or trait.

One commonly used measure of internal consistency reliability is Cronbach's alpha coefficient. Cronbach's alpha coefficient ranges from 0 to 1, with higher values indicating greater internal consistency reliability. A Cronbach's alpha coefficient of 0.7 or higher is generally considered acceptable for most purposes, while a coefficient of 0.8 or higher is considered good [11].

In the context of multilevel English proficiency tests, internal consistency reliability can be used to assess the extent to which the different levels of the test measure the same underlying construct of English language proficiency. This can help ensure that the test is measuring language proficiency consistently across different levels, and can provide evidence of the validity of the test.

Table 3 presents the Cronbach's alpha values for each version of the test, as well as the raw scores corresponding to the cut-scores identified by Rasch model. The calculations are done using ltm packet in R, a software environment for statistical computing and graphics [12].

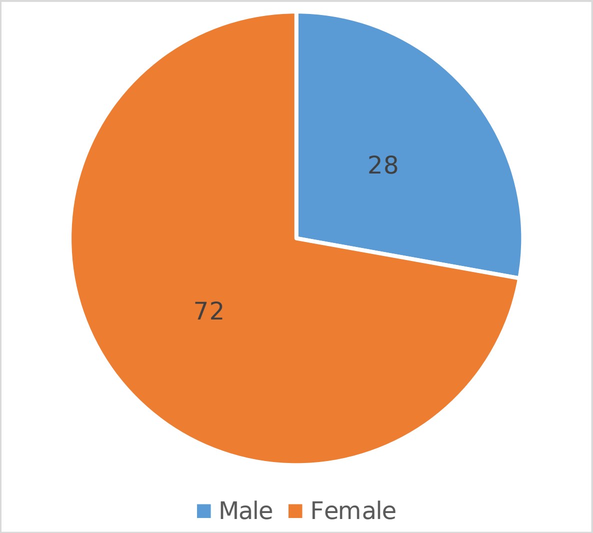
Version	**Listening**				**Reading**			
	Cronbach's alpha	**Passing score**			**Cronbach's alpha**	**Passing score**		
		B1	**B2**	**C1**		**B1**	**B2**	**C1**
22031	0.763	12	19	26	0.789	14	22	29
22032	0.812	12	20	27	0.843	10	18	27
22041	0.72	12	18	26	0.778	11	20	28
22042	0.84	9	16	25	0.832	11	19	28
22043	0.82	9	17	25	0.823	14	22	30
22051	0.81	11	19	28	0.796	8	16	25
22052	0.807	7	15	23	0.823	10	18	27
22053	0.723	11	20	28	0.766	9	18	27
22054	0.759	7	15	25	0.78	12	21	28
22061	0.804	8	16	25	0.823	11	19	28
22062	0.835	7	15	23	0.828	10	18	26
22063	0.839	10	18	27	0.817	7	15	24
22064	0.854	11	19	27	0.801	8	16	25

22065	0.829	8	15	24	0.829	11	19	28
22091	0.869	10	18	26	0.864	10	18	27
22092	0.858	10	18	26	0.873	11	20	28
22101	0.83	11	20	27	0.828	10	17	26
22102	0.858	11	19	27	0.787	10	18	26
22103	0.815	11	19	27	0.822	12	20	28
22104	0.841	13	22	30	0.834	12	20	29
22105	0.822	10	19	27	0.799	9	17	25
22111	0.755	11	18	25	0.757	10	17	25
22112	0.836	8	16	28	0.877	9	18	27
22113	0.805	6	13	23	0.881	9	18	27
22114	0.848	9	17	25	0.843	9	18	27
22115	0.885	10	18	26	0.856	7	16	25
22121	0.818	10	19	27	0.779	6	14	24
22122	0.832	9	17	26	0.853	9	18	27
22123	0.881	12	20	28	0.821	11	19	28
22124	0.85	10	18	27	0.887	12	20	28
22125	0.799	8	14	22	0.794	10	17	25
22126	0.802	10	18	26	0.787	8	16	25
22127	0.762	8	15	23	0.745	7	15	24
22128	0.847	7	14	23	0.819	11	19	27
Mean	**0.818**	**10**	**17**	**26**	**0.819**	**10**	**18**	**27**

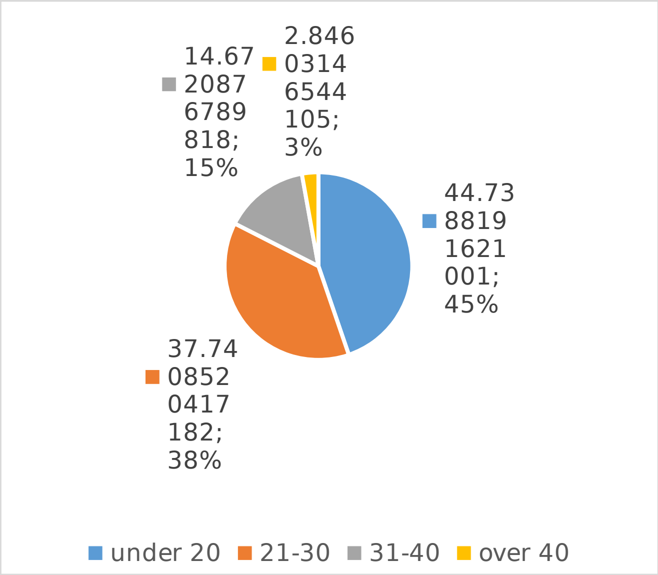
Table 3. Test performance statistics

4. Analysis of test taker performance

The demographic data of the test takers is shown in the following graphs. The data shows that every third test taker is a male participant. Most of the test takers are under 30 years old.



Graph 2. Test takers by gender (percentage)



Graph 3. Test takers by age group

The test taker performance for sessions conducted between March and December in 2022 is analysed across gender, age and regions. Overall performance results are shown in *Table 4.*

Level	Number of sertificates issued	Percentage
C1	3285	7,26
B2	21384	47,25
B1	15378	33,98

Below B1 (Fail)	5208	11,51

Table 4. Overall test taker performance

Gender	Below B1	B1	B2	C1
Male	17,19	37,94	39,22	5,65
Female	9,32	32,46	50,35	7,88

Table 5. Test taker performance by gender (percentage)

Region	Below B1	B1	B2	C1
Karakalpak Republic	12,65	27,88	50,84	8,63
Andijan	11,99	32,74	49,15	6,12
Namangan	11,51	34,88	46,98	6,62
Fergana	8,97	37,32	47,39	6,32
Bukhara	6,86	36,59	48,65	7,90
Samarkand	13,17	34,65	45,48	6,70
Navoiy	10,49	40,81	44,09	4,61
Jizzakh	11,57	30,60	49,79	8,05
Sirdarya	9,52	34,69	49,48	6,31
Kashkadarya	11,66	36,24	46,86	5,24
Surkhandarya	10,41	34,68	48,58	6,32
Khorezm	7,55	31,65	51,59	9,21
Tashkent	16,71	29,81	44,33	9,15

Table 6. Test taker performance by regions (percentage)

Age group	Below B1	B1	B2	C1
under 20	8,57	48,20	40,97	2,26
21-30	11,83	24,30	51,81	12,07
31-40	18,51	18,93	52,77	9,79
over 40	17,39	16,46	57,07	9,08

Table 7. Test taker performance by age group (percentage)

Gender	Listeni	Readi	Writi	Speaki	Over

	ng	ng	ng	ng	all score
Male	48,46	47,76	45,00	50,54	47,76
Female	50,25	49,94	50,86	53,77	51,23

Table 8. Mean scores by gender

Region	Listening	Reading	Writing	Speaking	Overall score
Karakalpak Republic	50,48	50,25	49,52	52,67	50,77
Andijan	49,44	49,38	48,93	53,73	50,35
Namangan	48,95	49,08	49,61	52,96	50,06
Fergana	49,42	48,99	50,05	53,33	50,48
Bukhara	50,07	50,07	51,33	53,75	51,31
Samarkand	49,06	48,45	49,24	52,14	49,70
Navoiy	48,86	47,75	48,42	52,08	49,30
Jizzakh	49,92	50,01	49,85	52,85	50,69
Sirdarya	49,61	49,25	50,99	53,41	50,80
Kashkadarya	49,64	48,88	47,98	52,01	49,61
Surkhandarya	49,22	49,56	49,78	52,99	50,34
Khorezm	51,72	50,40	51,67	53,87	51,97
Tashkent	50,02	49,64	46,71	52,41	49,51
Mean	**49,75**	**49,33**	**49,23**	**52,89**	**50,27**

Table 9. Mean scores by region

Age group	Listening	Reading	Writing	Speaking	Overall score
under 20	47,64	46,61	48,30	51,60	48,63
21-30	52,03	52,22	50,25	54,00	52,04
31-40	50,31	50,18	49,14	53,65	50,56
over 40	49,93	49,51	50,69	54,71	50,97

Table 10. Mean scores by age group

The tables above show that female test takers performed better than male participants. In terms of age groups, test takers between 21-30 years of age showed better performance than the rest of the age groups, while the youngest test takers received lower scores than the rest. Among regions, by far the best results were received by test takers from Khorezm and the Karakalpak Republic, while test takers from Kashkadarya and Navoiy regions received the lowest scores in comparison to other regions.

Conclusion

Multilevel English proficiency testing is an effective approach to evaluating language skills across multiple levels. By providing more accurate evaluations of language proficiency, better placement of students in language programs, and improved instruction and curriculum development, multilevel testing can help improve outcomes for students. Multilevel English proficiency testing is an evolving field, and there are many opportunities for future research and development.

Multilevel English proficiency test conducted by Assessment agency was designed using the latest accomplishments in classical and modern test theories. The analysis of test results show that the test results are reported on a single scale, which makes it comparable across different versions, and reliable enough to use for high stakes decisions in a local level.

Acknowledgements

The author would like to thank **Wobbe Zijlstra** of CITO, Netherlands for his exceptional support though online workshop training and **Mirshod Ermamatov** of Scientific and Educational Practical Center under the Agency for Assessment of Knowledge and Competences for his scientific advice on test analysis.

References

1. Brown, A. (2014). *Language assessment: Principles and classroom practices*. Pearson Education.

2. Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

3. Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

4. Baker, F.B. (2001). *The Basics of Item Response Theory, ERIC Clearinghouse on Assessment and Evaluation*. University of Maryland, College Park, MD.

5. Han, K.T. (2009). IRTEQ: Windows application that implements IRT scaling and equating [computer program]. *Applied Psychological Measurement*, 33(*6*), 491-493.

6. Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

7. Ermamatov, M. Dj., Abbosov, A.A, & Baratov, A.A. (2022). Test topshiriqlarini kalibrovkalash va qobiliyatlarni tenglashtirish. *Axborotnoma*, 3-4/2022, 4-15.

8. Abbosov, A.A. (2022). Yozish koʻnikmasini tekshirishda baholovchilar oʻrtasidagi ishonchlilik. *Axborotnoma*, 1-2/2022, 12-17.

9. Maris, G., Bechger, T., Koops, J., & Partchev, I. (2018). dexter: Data management and analysis of tests. URL: https://CRAN.R-project.org/package=dexter.

10. Kline, P. (1986). *A handbook of test construction: introduction to psychometric design*. London: Methuen.

11. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(*3*), 297-334. doi: 10.1007/BF02310555.

12. Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software*, 17, 1-15.

INGLIZ TILI BO‘YICHA KO‘P DARAJALI TEST TIZIMI TAHLILI

A.A. Abbosov

O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi huzuridagi Bilim va malakalarni baholash agentligi,
100084, Toshkent sh., Bog‘ishamol k., 12

Qisqacha mazmuni. Ko‘p darajali ingliz tilini bilish imtihoni test topshiruvchining ingliz tilini bilish darajasini turli darajalarda baholaydigan testdir. Ushbu turdagi testlar tobora ommalashib bormoqda, chunki u imtihon topshiruvchining til ko'nikmalarini yanada chuqurroq tushunish imkonini beradi. Ushbu maqolada ko'p darajali ingliz tilini bilish testining afzalliklari va qiyinchiliklari haqida ma'lumot berilgan hamda Bilimni baholash agentligi tomonidan o‘tkaziladigan ko‘p darajali test tizimining dastlabki natijalari tahlili keltirilgan. Test natijalar ko‘nikmalar kesimida, klassik va zamonaviy test nazariyalari asosida tahlil qilingan.

Kalit so‘zlar: Ko‘p darajali test tizimi, IRT, Rash modeli, qiyinlik darajasi, qobiliyat, standart ballar, ishonchlilik.