ANALYSIS OF PRELIMINARY RESULTS OF MULTILEVEL ENGLISH LANGUAGE PROFICIENCY TESTS

A.A. Abbosov

O‘zbekiston Respublikasi Vazirlar Mahkamasi huzuridagi Davlat test markazi,
100084, Toshkent sh., Bog‘ishamol k., 12

Abstract. The article discusses the results of the multilevel English proficiency tests conducted by Agency for Assessment of Knowledge and Competences. The test results are analyzed across different sections based on classical and modern test theories.

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

Based on the decree of the Cabinet of Ministers of the Republic of Uzbekistan No. 73 of February 16, 2022, a multilevel system of foreign language proficiency assessment was developed and put into practice. During the months of March-June 2022, 4 tests were organized and more than 17 thousand applicants participated in these tests.

Test results are analyzed based on classical and modern test theories. As results for listening and reading sections, latent ability scores based on the Rash model are presented in the form of standard scores. Cut scores are distributed as follows:

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

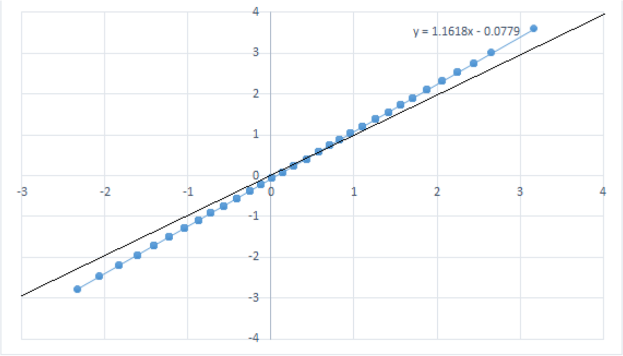
The results are as follows for the candidates who participated in March, April and May:

Level	Number of candidates	Per cent
C1	794	10,65
B2	3884	52,11
B1	2042	27,40
Below B1	457	6,13

The equating method is used to ensure the parallelism of test forms when calculating test results. In this case, the items used in the tests consist of non-repeating (unique) and overlapping test items:

Items\Versions	V1	V2	V3	V4
Overlapping	8	8		
	6		6	
	6			6
Unique	15	27	29	29
Total	35	35	35	35

The equating method allows to calculate the results of parallel tests on a single scale. Through this, not only several test versions used in one exam, but also the results of exams conducted at different times can be compared. The chart below compares the results of the tests conducted in April and May:



It can be seen that slope and intercept values of the linear function are very small. The error at both ends of the line does not significantly affect the test results. Also, the results are mainly recorded in the middle part, where the amount of error is small.

The reliability in the assessment of writing and speaking skills requiring human participation is constantly monitored. Pearson's correlation coefficient is used in reliability analysis.

Inter-rater reliability for Writing:

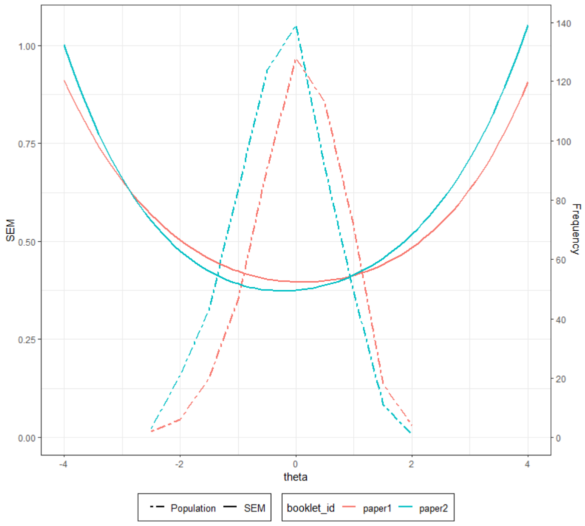
Criteria	March	April	May
Task achievement	0,824	0,794	0,841
Grammatical range and accuracy	0,821	0,800	0,854
Vocabulary range and appropriacy	0,825	0,798	0,859
Coherence and cohesion	0,827	0,808	0,848
Overall	0,876	0,857	0,899

Inter-rater reliability for Speaking:

Criteria	March	April	May
Discourse management	0,710	0,673	0,756
Grammatical range and accuracy	0,710	0,694	0,724
Vocabulary range and appropriacy	0,704	0,688	0,737
Pronunciation	0,638	0,632	0,681
Overall	0,785	0,768	0,816

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

In order to develop a multi-level test system, cooperation with CITO experts is underway. The validity and reliability of the test results are being studied based on the Rash model. Together with the CITO expert, the results of the test conducted in March were analyzed and it was noted that the reliability of the test results is high:



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 abilities. As can be seen from the graph, the results are mainly recorded in the interval where the error value is small.

The Department of Assessment of Foreign Language Proficiency is continuing to work on further improvement of the multilevel system.

ADABIYOTLAR

1. Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company.

2. Sarah Sharratt, Kembridj o‘quv seminari materiallari, 13-24 yanvar, 2021 yil.

3. Crocker L. and Algina J. (2008), Introduction to classical and modern test theory, Ohio: Cengage Learning.

4. Krathwohl, D.R., Bloom, B.S., & Masia, B.B. (1964). *Taxonomy of educational objectives: The classification of educational goals. Handbook II: Affective domain*. New York: David McKay Co.

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

A.A. Abbosov

State Test Center under the Cabinet of Ministers of Republic Uzbekistan,
Tashkent 100084, Bogishamol st. 12.

Qisqacha mazmuni. Ushbu maqolada 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.