UZBEKISTAN MINISTRY OF DEFENCE FOREIGN LANGUAGE APTITUDE TEST BATTERY: LEXICAL ANALOGIES SUBTEST ITEM QUALITY ASSESSMENT

Ilya Sergeevich Zverev
Partnership for Peace Training Center of the Armed Forces of the Republic of Uzbekistan, English language department head
zverev.elijah@gmail.com

ABSTRACT

Ministry of Defence of the Republic of Uzbekistan (Uzbekistan MoD) is the only governmental establishment of the nation providing its service members and employees with intensive foreign language training for the purpose of their participation in various events of international military and military technical cooperation. To this end, Uzbekistan MoD has put in place a number of mechanisms for the intensive foreign language training candidate selection, of which the central place is occupied by Foreign Language Aptitude Test Battery (UzMoD FLA TB). Since the results obtained by test takers in UzMoD FLA TB can be either beneficial or detrimental for their future careers, it is of utmost importance that the test items comprising it be of the highest quality possible. Within the framework of the present study, therefore, we conduct an assessment of the quality of the items comprising Subtest 1 of UzMoD FLA TB, i.e. Lexical Analogies Subtest through the prism of three basic parameters (item facility value, distractor efficiency and item discrimination index) based on the data from 137 service members of Uzbekistan MoD. We demonstrate that some items do not meet the quality requirements based on the parameters set and provide out recommendations as to how this particular situation can be rectified.

Keywords: Uzbekistan Ministry of Defense, intensive foreign language training, foreign language aptitude test battery, test item quality

INTRODUCTION

Intensive foreign language training (IFLT) that for the purposes of the present article is defined as “a foreign language teaching and learning within a certain period of time during which a student with zero to none foreign language learning experience gradually moves through a series of foreign language proficiency levels in order to achieve the one required by his/her superiors dedicating to this endeavor up to 8 hours daily” (Zverev, 2019, p. 139) demands that a certain candidate selection mechanism be put in place in order to guarantee the attainment of the results required within the time limits set.

To this end, Uzbekistan MoD FLA TB is used. In essence, it is a collection of four Subtests that are claimed to measure certain cognitive abilities presence of which in a candidate can be regarded as an indicator of his/her potential success in IFLT courses. The four Subtests in question are Lexical Analogies (Subtest 1), Shape Selection (Subtest 2), Linguistic Decoding (Subtest 3), and Narration Summary (Subtest 4).

Lexical Analogies Subtest is a collection of 30 multiple-choice items that very tenuously (see, for instance our discussion in (Zverev, 2019, p. 148) can be claimed to measure a candidate’s native language lexical proficiency alongside with his or her ability to establish logical connections of various types.

The principal aim pursued within the framework of the present article, therefore, is quality assessment of each of the items comprising Lexical Analogies Subtest of Uzbekistan MoD FLA TB for their subsequent improvement or replacement should such prove to be necessary.

LITERATURE REVIEW AND RESEARCH METHODOLOGY

Foreign language aptitude (FLA), a set of diverse cognitive skills or abilities generally viewed as conducive for foreign language training, is an attribute whose existence cannot be proven or disproven via direct observation (Zverev, 2019, 2020a, 2020b).

Consequently, there exist a number of theories and hypothesis as to what abilities comprise FLA in the first place. Among the most commonly named are “phonetic coding ability, grammatical sensitivity, inductive language learning ability, rote memory ability, grammar sensitivity, native language vocabulary range, native language skills, attentional control, working memory, language analysis ability, retrieval memory, perceptual speed, pattern recognition, etc” (Zverev, 2021, p. 1912).

Based on the results of even the most superficial analysis of all those theories and hypotheses, it can be concluded FLA is not regarded as a monolith, but rather as a multi-componential psychological construct, which is reflected in various measurement instrument developed and employed for its indirect assessment.

Uzbekistan MoD FLA TB Subtest 1 (Lexical Analogies) is one of the instruments utilized for the purposes of measurement of native language vocabulary range (postulated as a FLA component by Grigorenko (2002, p. 97)) and logical thinking ability (not traditionally included in any modern FLA conceptualization model). The Subtest, therefore, is language specific, and exists in two forms: one for Russian-as-a-native-language test takers and one for Uzbek-as-a-native-language test takers.

The subtest proper comprises 30 multiple-choice questions with identical structure: the stimulus material is a combination of two lexical items in the test taker's native language (Uzbek or Russian) with an underlying logical relationship of a certain kind (opposition, similarity, causation, etc). The test taker is to deduce the relationship expressed by the pair and select among the five variants provided that which expresses the same relationship.

The total amount of time allocated for Uzbekistan MoD FLA TB Subtest 1 is 8 minutes. Each correct response is worth one point and the test taker is not penalized for an incorrect response.

For the purposes of the present study, we analyzed the responses to Uzbekistan MoD FLA TB provided by 137 test takers (see ***Table 1***). They were tested within the time limit set in five separate rooms (up to 25 persons in each) with five invigilators present. The instructions for the test were provided first in Uzbek and then in Russian languages (if necessary). At the end of the test, the Response Matrices were collected by the invigilators and checked in a separate room by two designated officials. The results were subsequently analyzed by means of IBM SPSS Statistics 26 software package.

Table 1

Rank-Based Distribution of the Study Participants (N = 64)

Rank	*f*	Rel. *f*	C*f*	Percentile
Lieutenant Colonel	6	0,044	137	100,00
Major	8	0,058	131	95,62
Captain	22	0,161	123	89,78
Senior Lieutenant	31	0,226	101	73,72
Lieutenant	40	0,292	70	51,09
Sergeant	1	0,007	30	21,90
Junior Sergeant	6	0,044	29	21,17
Private	11	0,080	23	16,79

In order to assess the items comprising Uzbekistan MoD FLA TB Subtest 1, we chose three parameters: ***item facility value (IFV)***, ***distractor efficiency (DE)*** and ***item discrimination index (IDI)***. It behooves us to emphasize that none of them belongs to the aptitude testing area exclusively: all are applied in order to assess items comprising multiple-choice tests of various subjects.

IFV is the percentage of students who provided the correct answer to the test item. Generally, IFV is denoted as ***p***, which conveys the relative frequency with which the test taker answered the item correctly.

For instance, the *p* value for an item to which 55% of the test takers gave the correct answer, would be 0.55. The higher the item *p* value, the less trouble would the particular population sample representatives have dealing with it.

DE value is the degree of "attractiveness" alternative responses of a multiple-choice test item possess. This parameter is a numerical expression of the degree of usefulness of incorrect variants provided by the test developer (Gervais et al., 2015, pp. V–15). Thus, a particular distractor's attracting too large a number of test takers might be viewed as an indicator of its ambiguousness, whilst one not attracting anybody might be obviously incorrect and, consequently, having no importance in the testing process.

If there is a harmony between the object being measured by the test in general and a test item in particular, it makes sense to expect that those who achieve positive results on the test would answer the item correctly and vice versa. An item is considered to be a good one should it discriminate between the people scoring high on the test as a whole and those who score low. ***IDI*** is the quantified expression of the extent to which "the items separate the stronger test takers from the weaker ones in the positive or negative way" (Green, 2019, p. 21). The higher the IDI, the better the item.

If every test taker provides either the correct or the incorrect response to a particular test item, such an item should be removed or rewritten (IDI = 0.00). If every low-performing test taker answers the item correctly, whilst nobody in the upper group provides the correct response, the item is behaving in the direction opposite to that of the entire test and must be removed (IDI = -1.0). According to Urbina (Urbina, 2004, p. 231), "[f]or the vast majority of tests, discriminating power is the most basic quality that items must have in order to be included in a test".

Each of the three parameters being numerical, there was a need for determination of the threshold value that would separate what can be termed "acceptable items" from "unacceptable ones".

Taking into account that any FLA test is essentially an ability test that is employed in order to differentiate among individuals comprising a population sample based on a particular trait of interest taken to be normally distributed within the population, IFV value cannot be excessively narrow in order to account for both those with the highest level of the trait of interest and those with the lowest one. Consequently, "the p value of items should cluster around .50 (or 50%) to provide maximum differentiation among test takers" (Urbina, 2004, p. 230). What also follows is that extremely simple or complex items (with IFV close to 0% or to 100%) must not be included into aptitude/ability tests due to the fact of their failing to “differentiate among test takers and … [being] excess baggage” (Urbina, 2004, p. 230).

DE values of an item directly affect its IFV in at least two ways. First, the lower the number of distractors, the higher the chance of guess-response to an item. Second, the lower the quality of distractors, the easier it is to give the correct answer. Within a test, not only should the correct response be obvious to the test taker who possesses the trait being required to give that response, but also the distractors provided must look plausible enough to those lacking in such a trait. Therefore, DE “minimum threshold”, following Green (Green, 2019, p. 24), has been taken to be 7% and those distractors attracting fewer than that percentage of the test takers will be considered to be in need for alteration or replacement (depending on the exact DE value for each of them).

IDI value separating acceptable test items from unacceptable ones has been taken to be +0.3 and above (Green, 2019, p. 24).

Another parameter that we have considered was that of ***internal consistency*** defined as the degree of their being composed of items resulting in the same response patterns among individuals taking them. For example, two candidates obtaining the same score on a test should demonstrate the same pattern of correct and incorrect responses. As emphasized by Gervais (2015, pp. c–3), “the more homogeneous the domain tested, the higher the internal consistency”. Internal consistency can be reported by means of either Kuder Richardson 21 (KR21) or the Cronbach’s Alpha coefficient. The latter being the most commonly used one in modern testing practices and research, it has been chosen as an assessment parameter in our research. The cutt-off point for the Cronbach’s Alpha coefficient has been taken to be 0.80, as per suggestion by Gervais (2015, pp. c–6).

RESULTS

The results of the analysis of items comprising Subtest 1 are shown in Table 2.The Cronbach Alpha coefficient value calculated based on the FT-II data was 0.875, which is above the established cut-off point. Therefore, we can confirm that Subtest 1 does have internal consistency. There are, however, a number of issues pertaining to the test items in terms of the parameters selected for the assessment purposes that need to be addressed.

Table 2

Uzbekistan MoD FLA TB Subtest 1 Item Analyses (Cronbach's Alpha = 0.875)

Item	IFV[a]	Distractor Efficiency		CITC[b]	CAID[c]	Item	IFV[a]	Distractor Efficiency		CITC[b]	CAID[c]
1.	85.4%	а = 8.0% б = 1.5% в = 0.7%	***г*** = 85.4% д = 2.9 х[d] = 1.5	0.063	0.877	**16.**	67.2%	***а*** = 67.2% б = 5.1% в = 3.6%	г = 10.9% д = 12.4% х = 0.7%	0.132	0.878
2.	81.8%	а = 4.4% б = 2.2% в = 10.2%	***г*** = 81.8% д = 1.5% х = 0%	0.234	0.875	**17.**	51.8%	а = 21.2% ***б*** = 51.8% в = 2.9%	г = 8.0% д = 11.7% х = 4.4%	0.537	0.868
3.	66.4%	а = 0.7% б = 6.6% в = 24.1%	***г*** = 66.4% д = 0% х = 2.2%	0.348	0.872	**18.**	77.4%	а = 6.6% б = 1.5% в = 2.2%	г = 11.7% ***д*** = 77.4% х = 0.7%	0.372	0.872
4.	80.3%	***а*** = 80.3% б = 4.4% в = 2.9%	г = 7.3% д = 3.6% х = 1.5%	0.409	0.871	**19.**	21.9%	а = 44.5% б =21.9% в = 2.9%	***г*** = 21.9% д = 7.3% х = 1.5%	0.200	0.876
5.	46.7%	а = 18.2% б = 19.0% ***в*** =46.7%	г = 14.6% д = 1.5% х = 0%	0.352	0.872	**20.**	58.4%	а = 7.3% б = 4.4% ***в*** = 58.4%	г = 7.3% д = 21.9% х = 0.7%	0.544	0.867
6.	69.3%	а = 1.5% б = 14.6% в = 8.8%	г = 5.8% ***д*** = 69.3% х = 0%	0.658	0.865	**21.**	75.2%	***а*** = 75.0% б = 5.1% в = 5.1%	г = 8.8% д = 5.1% х = 0.7%	0.585	0.867
7.	68.6%	***а*** = 68.6% б = 16.1% в = 5.1%	г = 5.1% д = 4.4% х = 0.7%	0.394	0.871	**22.**	46.7%	а = 30.7% ***б*** =46.7% в =6.6%	г =3.6% д = 10.2% х = 2.2%	0.354	0.872
8.	72.3%	***а*** = 72.3% б = 5.1% в = 3.6%	г = 2.2% д = 16.8% х = 0%	0.614	0.866	**23.**	89.1%	а = 3.6% ***б*** =89.1% в =2.9%	г =2.2% д = 2.2% х = 0%	0.342	0.873
9.	70.8%	а = 8.8% ***б*** = 70.8% в = 9.5%	г = 5.8% д = 5.1% х = 0%	0.647	0.865	**24.**	73.0%	а = 3.6% ***б*** = 73.0% в = 6.6%	г = 5.1% д = 6.6% х = 5.1%	0.509	0.869
10.	78.8%	а = 8.8% б = 6.6% в = 4.4%	***г*** = 78.8% д = 1.5% х = 0%	0.556	0.868	**25.**	76.6%	а = 2.9% б = 7.3% ***в*** = 76.6%	г = 2.2% д = 7.3% х = 3.6%	0.404	0.871
11.	70.1%	а = 10.2% б = 15.3% в = 2.2%	г = 1.5% ***д*** = 70.1% х = 0.7%	0.580	0.867	**26.**	45.3%	а = 18.2% б = 8.0% в = 10.2%	г = 14.6% ***д*** = 45.3% х = 3.6	0.526	0.868
12.	80.3%	а = 0.7% ***б*** =80.3% в =2.9%	г = 6.6% д = 9.5% х = 0%	0.567	0.868	**27.**	34.3%	а = 6.6% ***б*** = 34.3% в = 5.1%	г =42.3% д = 5.1% х = 6.6%	0.185	0.877
13.	47.4%	а = 2.9% б = 8.0% ***в*** = 47.4%	г = 35.0% д = 5.8% х = 0.7%	0.373	0.872	**28.**	51.1%	***а*** = 51.1% б = 6.6% в = 30.7%	г =3.6% д = 3.6% х =4.4%	0.336	0.873
14.	78.8%	а = 11.7% б = 5.8% в = 2.2%	г = 1.5% ***д*** = 78.8% х = 0%	0.615	0.866	**29.**	81.0%	***а*** = 81.0% б = 4.4% в =7.3%	г = 2.9% д = 0% х = 4.4%	0.477	0.870
15.	57.7%	***а*** = 57.7% б = 38.7% в = 0.7%	г = 0.7% д = 2.2% х = 0%	0.015	0.881	**30.**	51.1%	а = 29.2% б = 2.9% в = 4.4%	***г*** =51.1% д = 7.3% х = 5.1%	0.386	0.872

Note. [a] Item Facility Value. [b] Corrected Item-Total Correlation (Item Discrimination Index). [c] Cronbach's Alpha if Item Deleted [d] The percentage of test takers who did not provide any answer to the test item. The keys to each question are indicated in ***italicized bold text***.

DISCUSSION

First, 80% of the Subtest 1 items had IFV above 50% with the lowest being that of item #19 (21.9%) and the highest – that of item#23 (89.1%). Subtest 1 item facility values are negatively skewed (-.808), which results in “insufficient ceiling”, i.e. in the situation when individuals taking the test do not find the items challenging enough (Urbina, 2004, p. 230).

Second, the results of DE analysis indicate that in 50% of the Subtest 1 items there was only one distractor that managed to attract more than 7% of the responses provided by the test takers. Since a test item has 4 distractors plus 1 correct response, we would argue that half of the items comprising Subtest 1 have only one distractor that is not obviously wrong, which increases the chances of giving the correct response to such items to 50%.

In some test items, a distractor attracted a proportion of responses comparable to that attracted by the key option (47.4% of the test takers gave the correct response to item #13, yet 35% of them selected one of the distractors provided). Conversely, the distractor managed to act better than the key option in item #27 (42.3% of the test takers chose a distractor as their response).

Third, DE values immediately affect items’ IDIs. The minimum Subtest 1 IDI was that of item #15 (0.015) and the maximum was that of item #6 (0.658). The mean IDI for the 30 items did not exceed 0.41 and 20% of the items (#1, 2, 15, 16, 19 and 27) had IDI below 0.3. Consequently, a fifth of all the items comprising the Subtest 1 do not actually possess the required discrimination power in order to separate low- from high-performers in the test.

CONCLUSION

The results of the item quality analyses conducted strongly suggest the need to modify a major number of test items comprising Uzbekistan MoD FLA TB. Those test items need to be reviewed in order to deal with their excessive facility arising due to the poor quality of distractors and resulting in insufficient power to discriminate between weaker and stronger test takers

REFERENCES

1. Gervais, A. (2015). *The Concepts of Reliability and Validity in Military Training and Education* (p. 20).

2. Gervais, A., Schwarz, M., Seinhorst, G., Rey, C., & Hezog, M. (2015). *Bureau for International Language Coordination Language Testing Seminar Materials Pack*. NATO BILC.

3. Green, R. (2019). Item Analysis in Language Assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment. Volume I* (pp. 15–30). Routledge.

4. Grigorenko, E. L. (2002). Foreign language acquisition and language-based learning disabilities. In P. Robinson (Ed.), *Individual Differences and Instructed Language Learning* (pp. 95–112). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.2.07gri

5. Urbina, S. (2004). *Essentials of Psychological Testing* (A. S. Kaufman & N. L. Kaufman (eds.)). John Wiley & Sons, Inc. https://doi.org/10.1037/h0053347

6. Zverev, I. (2019). Uzbekistan MoD Foreign Language Aptitude Test: A Critical Evaluation. *Filologiya Masalalari*, *29*(2), 138–151.

7. Zverev, I. (2020a). Uzbekistan MoD Foreign Language Aptitude Test Battery Predictive Power Analysis. *European Journal of Research and Reflection in Educational Sciences*, *8*(7), 35–44. https://doi.org/10.36078/987654354

8. Zverev, I. (2021). Revisiting Predictive Power Analyses of Uzbekistan MoD Foreign Language Aptitude Test in Terms of Intensive English Language Training Success. *Academic Research in Educational Sciences*, *2*(4), 1910–1917. https://doi.org/10.24411/2181-1385-2021-00818

9. Zverev, I. (2020b). Computer-based Foreign Language Learning Deficienty Diagnostics Using LLAMA Test Battery. *Chet Tillarni O'qitishda Xorijiy Tajribalardan Foydalanish*, 72–74.