 DOI: 10.3849/2336-2995.32.2023.01.118-134

Peer-reviewed

One Standard, Different Approaches: Language Assessment in the Military Context. A Visegrad countries analysis.

Jeden standard, různé přístupy: Hodnocení jazykové způsobilosti ve vojenském kontextu. Analýza přístupů v zemích V4.

Ivana Mrozková, Mária Šikolová

Abstract: The article describes how one standard set for language testing is perceived in Visegrad countries and how it is transformed into designing their language proficiency tests. It focuses on analyzing the methods of how the four countries use the identical descriptors to develop their own distinct tests. In their comparison and analysis, the authors concentrate on the test format, testing methods, tester training, and assessment techniques. The gathered data have shown some similar approaches to test design and administration, however, some differences in certain aspects were also observed.

Abstrakt: Článek se zabývá způsobem, jak je jeden standard stanovený pro jazykové testování vnímaný v zemích V4 a jak jej jednotlivé země transformují do tvorby jejich zkoušek jazykové způsobilosti. Zaměřuje se na analýzu metod, kterými země V4 používají jednotné deskriptory k tvorbě svých národních zkoušek. Autorky se ve své analýze a komparaci soustřeďují zejména na formát testů, testovací metody, školení testerů a způsoby hodnocení. Na základě získaných dat lze říci, že v přístupech k tvorbě testů a jejich administraci existuje ve zkoumaných zemích hodně podobností, nicméně v některých aspektech země přistupují k hodnocení jazykové způsobilosti odlišně.

Key words: Language Test Analysis; Descriptors of Language Skills; Test Format; Language Assessment; NATO STANAG 6001.

Klíčová slova: Analýza jazykových testů; deskriptory řečových dovedností; formát testu; hodnocení jazykové způsobilosti; NATO STANAG 6001.

INTRODUCTION

The importance of language assessment is growing in the current era of standardization and globalization. Speaking a common language has become essential in many spheres of activity. To tackle the challenges humans face from fighting diseases and the consequences of natural disasters to fighting the consequences of human behavior, a unified approach to language standardization is inevitable.

The military context is no exception; good communication is vital for co-operation in military operations. Within the framework of NATO, the language of inter-operational communication is English; therefore, to ensure that every individual part of the joint forces communicates on a certain level of English became crucial almost from the very beginning of the Alliance. The creation of one standard to describe language proficiency levels was necessary and had to be approached on the NATO top command level.

NATO, with help of its advisory body, Bureau for International Language Coordination (BILC) issued NATO Standardization Agreement 6001 (NATO STANAG 6001) in 1976, which was subsequently ratified by NATO member states. Currently, the second version of Edition 5, issued in 2016, is valid. NATO STANAG 6001 serves as the agreed-upon NATO standard for language curriculum design, test development, and for recording and reporting Standardized Language Profiles (SLPs), defining 6 basic proficiency levels: 0 (no proficiency), 1 (survival), 2 (functional), 3 (professional), 4 (expert), and 5 (highly-articulate native). According to NATO STANAG 6001, four basic language skills are assessed separately, resulting in SLP describing language proficiency of individuals in listening comprehension, speaking, reading comprehension and writing (in this order).

Unlike other assessment systems which are characterized by the same assessment tool being used indiscriminately in every country in which the system is utilized, within the NATO STANAG 6001 framework, each signatory country develops their own tests reflecting their unique conditions and procedures, provided that they abide by the descriptors of NATO STANAG 6001. In other words, NATO STANAG 6001 has established a norm which should be applied in all NATO countries, therefore, the results of the examination reported in the SLP are valid in all NATO countries. Moreover, it is important to note that the results of these examinations are commonly of high importance and can have a significant impact on the careers of professional soldiers.

To better understand how one standard can function in different countries and under different conditions, the authors have decided to gather data concerning essential parts of language testing systems in the Visegrad countries (Czech Republic, Hungary, Poland, Slovakia). The rationale behind this choice was the fact that these countries used to be parts of the former Soviet Block and hence, they share a similar cultural and political background: in the 1990s, all underwent profound changes in terms of the political system, a transition which was consequently reflected in all segments of life, including the military.

This survey aims at describing, comparing, and analyzing the language assessment systems developed in the above-mentioned countries following the NATO STANAG 6001 descriptors. In addition, it discusses the advantages and disadvantages of the used tools and techniques and draws conclusions about the unique language assessment concept

of having one standard implemented through the different test designs and testing techniques used.

1 RESULTS

The authors of the article wanted to find out how one common standard (NATO STANAG 6001) can be approached by different national testing boards. Thus, they have decided to compare relevant areas of language testing in Poland, Hungary, and the Slovak and the Czech Republics. Testing specialists from the afore-mentioned countries provided the authors with data describing the testing procedures and techniques used in their respective countries at the English exam in accordance with (further i.a.w.) NATO STANAG 6001 in 2019 as the data from later two years were influenced by the COVID 19 pandemic.

1.1 Tests Characteristics

The approaches to some of the areas of applying NATO STANAG 6001 standards to testing were very similar in all countries included in the study, e.g. all countries conduct moderation sessions when creating individual test items, as well as all testing teams pre-test their tests prior to its use.

Further, another very similar approach is being applied in setting the cut score. In two countries, it is arbitrarily stated (70% and 60%); in one country it is between 65 and 70% depending on pre-testing results, and in one country the cut score is set in dependence on the results of pretesting.

As far as the required parameters of classical test theory are concerned, the acceptable facility value of test items ranges mostly from 30 to 70%; in one country, different facility value is required for levels 1-2 (70-90%) and 2-3 (50%-70%). Desirable discrimination index in three countries is 0.30+, in one country only it is 0.10+. The acceptable value of Cronbach's Alpha is around 0.70 - 0. 80.

As mentioned above, NATO STANAG 6001 distinguishes 6 different levels of proficiency in all four skills. In addition, the countries may decide to use plus levels, e.g. L2+ which means that the candidate at this level meets most of the requirements of the higher level (L3), but does not produce the language at the desired level consistently. Our survey has shown that three out of four countries use the plus levels.

Language of test instructions was also surveyed; two countries use the target language in the instructions; the other two use it as well, but with the exception of level 1 examinations where they use the native language.

1.2 Testing techniques in testing individual skills

The types of testing techniques, as well as the time limits for particular skills and proficiency levels vary broadly in the countries under the survey.

1.3 Testing listening comprehension

In testing listening comprehension, testing techniques used as well as testing conditions vary considerably. In two countries, listening tests are done with headphones in laboratories. In one country, there is no lab at all, in another one, there are no headphones. In three countries, all recordings are played twice; in one country, the recordings for levels 2 and 3 are played once only; for level 1, the repetition of a recording depends on how long and difficult the text is.

Only one country uses exclusively multiple-choice questions (MCQs) for testing listening at all levels. Otherwise, different combinations of testing methods are used, such as MCQs, short answers and true – false statements; MCQs, open answers and gap-filling; MCQs, constructed responses, and short answers. As for the time limit, in the tests for level 1, it varies from 20 to 35 minutes, in the tests for level 2 the range is even bigger from 25 to 45 minutes; the biggest difference is in fact in the tests for level 3 ranging from 30 to 60 minutes. Another interesting point concerning the length of the listening subtest is, how it differs in each country level-wise. While in two countries the length of the listening part is the same for levels 1 and 2, the difference in the time limit between these two levels (1, 2) and level 3 in one of these countries is 10 minutes, in the other one it is doubled (from 30 to 60 minutes). In the other two countries, the time limit increases slightly as the levels are higher (25 – 30 – 35; 25 – 45 – 50).

Table 1: Comparison of testing techniques and time limits for testing listening comprehension

Country	L1 task types	L1 time limits	L2 task types	L2 time limits	L3 task types	L3 time limits
Country 1	MCQ, short response, true - false	35 minutes	MCQ, short response, true - false	35 minutes	MCQ, short response	45 minutes
Country 2	MCQ, open answers, filling the gaps	20-25 minutes	MCQ, open answers, filling the gaps	25-30 minutes	MCQ, open answers, filling the gaps	30-35 minutes
Country 3	MCQ, constructed response, short answer	30 minutes	MCQ, constructed response, short answer	30 minutes	MCQ, constructed response, short answer	60 minutes
Country 4	MCQ	25 minutes	MCQ	45 minutes	MCQ	50 minutes

Tables:

Not to identify the individual countries, the authors decided to allocate each country a number. The same numbers always refer to the same countries.

1.4 Testing reading comprehension

The variety of testing techniques in assessing reading skills is similar to the one used in testing listening skills. One country uses exclusively MCQs, while in the rest of the countries, other methods are used in combination with MCQs, such as true – false statements, matching, open answers, gap filling, constructed responses, and short answers.

The differences in the time limits for reading comprehension subtest are also quite considerable. Nonetheless, in all countries more time is allocated to measuring the proficiency level of reading skills than listening skills. While in two countries the time limit for reading skills at all levels is considerably longer if compared with listening skills (e.g. for listening L1 approx. 20-25 minutes, L2 approx. 25-30 minutes, and L3 approx. 30-35 minutes; for reading L1 it is 80 minutes, L2 it is 90 minutes, and L3 it is 90 minutes); in the other two countries there are just slight differences (e.g. for listening L1 approx. 25 minutes, L1-2 approx. 45 minutes, and L2-3 approx. 50 minutes; for reading L1 it is 30 minutes, L1-2 it is 45 minutes, and L2-3 it is 60-65 minutes).

Table 2: Comparison of testing techniques and time limits for testing reading comprehension

Country	L1 task types	L1 time limits	L2 task types	L2 time limits	L3 task types	L3 time limits
Country 1	MCQ, true –false, matching	70 minutes	MCQ, true –false, matching	70 minutes	MCQ	65 minutes
Country 2	MCQ, open answers, filling the gaps	80 minutes	MCQ, open answers, filling the gaps	90 minutes	MCQ, open answers, filling the gaps	90 minutes
Country 3	MCQ, constructed response, short answer	40 minutes	MCQ, constructed response, short answer	50 minutes	MCQ, constructed response, short answer	70 minutes
Country 4	MCQ	30 minutes	MCQ	45 minutes	MCQ	60-65 minutes

1.5 Testing speaking

As far as assessing speaking skills is concerned, there are also different approaches in different countries both in terms of testing techniques and time limits. In all countries, the speaking part of the examination is recorded. The most frequently used testing method is the role play, which is a part of the examination of all countries for levels 1 and 2. Two countries use picture description, however, not for the same levels – in one country it is used in level 1 exam, in the other one in level 2 exam. Only one country uses the technique called information gathering task, which is focused on integrated skills including asking questions, note-taking and reporting. In one country, at levels 1 and 2, both general and military conversations are part of the examination.

Not surprisingly, the broadest variety of elicitation techniques is used while testing the highest level, level 3. In one country, two candidates are examined at the same time - there are two presentations (briefings) at this level given by each candidate, both

followed by questions from the other candidate, discussion between these two candidates based on selected topics and the exam is finished by discussion prompted by the examiner. In other countries, the candidates are examined individually. Other countries do not use the format of presentations in level 3 examinations; the tasks used at this level are conversation, discussion, debate, interview, summary, expressing opinion on a given statement, hypothesizing, expressing opinion. Even though the names of the tasks look different at the first sight, however, the ways how the oral examination is conducted are quite similar.

The time limit set for measuring speaking proficiency at different levels varies broadly both within one country among the levels and among countries as well. Also, some of the countries stated the time limits in a range, e.g. 10-15 minutes. Probably, it depends on the candidates and the level of language they produce, i.e. whether they produce a rateable sample in a shorter time and thus it is possible to decide on the level easily, or whether a longer time is needed to measure the proficiency level appropriately. In most countries, the time limit for level 1 ranges between 10 and 15; only in one country it is between 20 and 25 minutes. As for level 2, the time limits are from 13 to 25 minutes, while for level 3 they range from 20 to 45 minutes. One country has time limits 12, 13 and 25 for levels 1, 2 and 3 respectively; one country has the time limit for level 1 from 20 to 25 minutes and for the other two levels the time limits are the same 25 – 30 minutes. In another country, the differences among the levels are even bigger from 10 – 15 minutes for level 1 through 15-25 minutes for level 2 and finally from 25 to 45 minutes for level 3. In the last country, the time limit rises with higher levels – 10, 15 and 20-25 minutes for levels 1-3.

Table 3: Comparison of testing techniques and time limits for testing speaking

Country	L1 task types	L1 time limits	L2 task types	L2 time limits	L3 task types	L3 time limits
Country 1	Warm-up Questions from an examiner Role-play with an examiner	12 minutes	Warm-up Role-play with an examiner Questions from an examiner	13 minutes	(in pairs) Presentation 1 (briefing) Questions from partner Presentation 2 (briefing) Questions from partner Discussion with a partner based on a selected topic Discussion prompted by an examiner	25 minutes
Country 2	General conversation Military conversation Picture description Role-play	20-25 minutes	General conversation Military conversation Discussion of some current military or military-political news Role play	25-30 minutes	Military conversation Discussion of a general topic Summary of a short military article and discussion of its topic	25-30 minutes

Country 3	Interview led by an interlocutor Situational role play/dialogue Describing, comparing and contrasting pictures	10-15 minutes	Interview led by an interlocutor Situational role play/dialogue Describing, comparing and contrasting pictures	15-25 minutes	Interview led by an interlocutor Expressing opinion on a quote Hypothesizing using a picture	25-45 minutes
Country 4	Introduction Role play	10 minutes	Introduction Role-play with complication Information gathering task	15 minutes	Introduction Interview with current news Debate	20-25 minutes

1.6 Testing writing

The same variation as while assessing the speaking skills can be witnessed in writing skills assessment. Countries differ in types of tasks the candidates are required to complete as well as in word and time limits set for the different levels.

Regarding the number of tasks and time limits for assessing level 1, only one country has one task only with the required number of 100 words. In the other countries, 2 tasks are administered at level 1 with no word limit set in one country; in the two others, the word limit differs (120/120 and 25-30/100-120) respectively. What is interesting is also setting the time limits which vary from 20 minutes through 30 and 60 up to the longest one of 80 minutes. The task types at this level also differ; the most frequent ones are short letters, invitations, and messages; the other types are e.g. instructions, a composition on a general topic, a letter to a friend on a military-themed topic, a form completion, a postcard, a reply to a note/message.

Table 4: Comparison of number of tasks, words and time limits for level 1 writing examination

Country	Number of tasks	Word limit	Time limit
Country 1	2	none	60 minutes
Country 2	2	240 words	80 minutes
Country 3	2	125 – 150 words	30 minutes
Country 4	1	100 words	20 minutes

To assess level 2 candidates, all countries use two different tasks with different time limits. In one country only, the recommended number of words is the same for both tasks (150-200). In two countries, the required number of words for the first task is lower than for the second one (80-100 and 150-200; 70 and 150). In the last country, however, the number of words for the first task (180) is higher than for the second one (150).

In task types, correspondence (formal/semiformal/informal) prevails. Similarly, the time limits set for writing subtests also differ; they are set in the range of 40, through 60 and 75 up to 90. Military topics are mentioned in two countries (report on a military

topic; a memo or a report on a military topic). In two countries in which the first task has a lower required number of words than the second one, the first task requires lower level skills and is similar to a level 1 task mentioned above (a note, a message, a reply to a note/message; a message, an invitation, a short personal letter).

Table 5: Comparison of number of tasks, words and time limits for level 2 writing examination

Country	Number of tasks	Word limit	Time limit
Country 1	2	300-400 words	75 minutes
Country 2	2	330 words	90 minutes
Country 3	2	230-300 words	60 minutes
Country 4	2	220 words	40 minutes

As far as the format of the level 3 examinations is concerned, all countries require the candidates to write 2 different tasks. The task types include formal/informal/semiformal correspondence; memo; report on social/military topic; an essay/a composition; magazine/ newspaper article. In one country only, the candidates are given an opportunity to choose either a military or non-military topic of composition. In all countries, the candidates are required to write two tasks with a different number of words, while one task is shorter and easier, the other one is longer and more demanding (120/200; 220/180; 180-200/250-300; 150/300). The time limits allocated to writing at this level are very similar (80-90 minutes).

Table 6: Comparison of number of tasks, words and time limits for level 3 writing examination

Country	Number of tasks	Number of words	Time limit
Country 1	2	120/200	80 minutes
Country 2	2	220/180	90 minutes
Country 3	2	180-200/250-300	80 minutes
Country 4	2	150/300	90 minutes

In terms of the time limits for writing in different countries, the most similar time is allocated for level 3 writing (80, 80, 90, 90 minutes). For level 2 writings, the differences are greater ranging from 40 through 60 and 75 up to 90 minutes. Nevertheless, the biggest differences can be observed among time limits for writing at level 1 (20, 30, 60, and 80 min.).

1.7 Testing boards

The number of testers reflects the needs for testing in particular countries. In one country, there are seven testers – members of Central Examinations Board with other testers from Military universities and Armed forces whose number is unknown. The next country has seven permanent positions for testing, five of them being testing specialists, the other two being administrators (not language specialists), as well as 40 contracted

testers and 10 contracted invigilators. In other two countries, the number of testers is 6 and 12 respectively. In the country with 12 permanent positions for testers, three other methodologists examine on regular basis and teachers take part in testing occasionally (when needed).

As for the qualifications of the testers, all countries require the testers to be university graduates (Linguistics/Philology/Pedagogy) with either teaching practice only or teaching practice in the military, or even teaching practice for NATO STANAG 6001. In one country only, the other pre-requisite for being a tester is to acquire NATO STANAG 6001 SLP 3333 minimum.

In all countries, testers have to undergo some training before they become members of the testing team and after that, there are regular calibration/norming sessions (once or twice a year). As mentioned above, BILC organizes various events whose purpose is to present best practices in language testing and harmonize the approaches in the countries which use NATO STANAG 6001.

The speaking examination is approached in different countries differently in terms of the number of testers, as well as their roles. In two countries, at the examinations for all levels, there are two testers, functioning as both interlocutors and raters. One country has two testers (functioning as both interlocutors and raters) for levels 1 and 2, while three testers for level 3. The last country has for all levels one interlocutor and two raters.

In assessing speaking and writing, identical methodology is applied in all four countries. Each writing is assessed by two raters independently and in case of disagreement, the third one makes the decision.

1.8 Testing organization and success rate

The testing institution is part of a university in two countries (University of Defense, University of Public Service). In one country, testing team is a part of the Language Institute under the command of the General Staff. In one country, testing of levels 1 and 2 is decentralized (the military universities and the armed forces can organize the examinations and they produce their own exams for levels 1-2); testing of level 3 is centralized and it is conducted by Central Examinations Board which is a part of Armed Forces School of Languages. The Central Examinations Board provides the other testing teams with exam materials for end-of-course sessions twice a year (two level 1 exams and two level 2 exams a year), as well as they train their testers in marking and test construction.

The number of test takers per year ranges from approximately 440 up to 4300. The following table illustrates the number of testers and test takers per year (figures from 2019).

Table 7: Comparison of the number of testers and average number of test takers per year

Country	Number of testers	Approx. number of test takers per year
Country 1	7 + other unknown number	4300 (tested by 7 testers)
Country 2	5+ 2 administrators (not language specialists), 40 contracted testers and 10 contracted invigilators	1000
Country 3	6	440
Country 4	12	4000

As far as the frequency of testing is concerned, it also varies – four, five, and ten times per year; in one country, the testing is conducted throughout the year with the exception of August, more-or-less four times per week.

The validity of examination certificate is unlimited in all four countries; however, in one country, if military are sent abroad to a mission, the certificate must be renewed if it is older than 3 years.

The subtests of individual skills can be retaken in three countries under specified conditions which differ; in one country, the whole examination must be retaken and paid for.

2 DISCUSSION

As mentioned above, in all researched countries the tests reflect the requirements set by NATO STANAG 6001 descriptors; however, assessment practice itself differs to a various degree in the countries under survey.

The procedure of test design and creation is similar in all countries under survey and is in compliance with BILC recommendations. Each test item undergoes several moderation sessions where it is investigated whether it fits the requirements set by the level descriptors and rating categories. Moderation is necessary to ensure that every item used in assessment contributes to the fair, valid, and reliable outcome of the assessment[1]. It also promotes consistency in assessment, together with pre-testing and constant test revision.

In all countries entering the study, pretesting is required which is in line with most testing specialists[1] who agree that in high-stakes tests pretesting is necessary to create tests that are truly relevant to distinguish between the levels set by language proficiency descriptors.

1 HUGHES, Arthur. *Testing for Language Teachers*. 2nd ed. Cambridge: Cambridge University Press, 2013. ISBN 9783125338760; BACHMAN, Lyle F. a Adrian S. PALMER. *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press, 1996. Oxford applied linguistics. ISBN 0194371484; MCNAMARA, T. F. Language testing. Oxford: Oxford University Press, 2008. Oxford introductions to language study. ISBN 0194372227.

2.1 Testing receptive skills

While there is consensus in moderation and pretesting stage of the test creation, test formats and testing techniques differ in each country. For testing receptive skills, similar techniques are used, with the exception of matching, which is used for testing reading comprehension only. Otherwise, one country uses exclusively MCQs for testing both reading and listening comprehension. In the other three countries, various methods are used in combination with MCQs, such as true – false statements, matching (for testing reading comprehension skill only), open answers, gap filling, constructed responses, and short answers.

The specialized literature does not recommend any universal testing techniques for testing receptive skills; nevertheless, quite a lot of specialists deal with different testing techniques highlighting their advantages and drawbacks.

Fixed response techniques, such as MCQs, true - false statements, and matching, are commonly used in tests because of their efficiency: the scoring is fast, economical and reliable [2]. As candidates choose between presented alternatives rather than constructing a response themselves [3], they are able to respond to more test items in a given period of time and they can get more "fresh starts" which contributes to higher reliability. Assessing fixed response items is fully objective and can be done by a non-specialist/ administrative workers or computer. Another important feature is that they "allow the testing of receptive skills without requiring the test taker to produce written or spoken language" [4]. If production, either spoken or written, is also required in testing receptive skills, the questions arise concerning scoring the answers – which answer is considered to be correct – the one which shows understanding, however might contain mistakes (grammar, syntactic, spelling…), or only the one which is correct in terms of meaning, as well as language used.

For all above-mentioned advantages, MCQs are a favorite assessment technique used by almost all assessment systems. However, apart from the apparent advantages of MCQs, there are also difficulties connected with this test technique: As it tests recognition knowledge only, using it can result in an inaccurate picture of the candidates' ability[5], test scores can be heavily influenced by guessing[6], and by the risk of cheating. Moreover, it is extremely difficult to write successful/functional MCQs items with the

2 HUGHES, Arthur. *Testing for Language Teachers*. 2nd ed. Cambridge: Cambridge University Press, 2013. ISBN 9783125338760.

3 MCNAMARA, T. F. Language testing. Oxford: Oxford University Press, 2008. Oxford introductions to language study. ISBN 0194372227.

4 HUGHES, Arthur. *Testing for Language Teachers*. 2nd ed. Cambridge: Cambridge University Press, 2013. ISBN 9783125338760.

5 HUGHES, Arthur. *Testing for Language Teachers*. 2nd ed. Cambridge: Cambridge University Press, 2013. ISBN 9783125338760.

6 ANDRICH, David a Ida MARAIS. Controlling Bias in Both Constructed Response and Multiple-Choice Items When Analyzed With the Dichotomous Rasch Model. *Journal of Educational Measurement* [online]. 2018, **55**(2), 281-307 [cit. 2022-07-20]. ISSN 00220655. Dostupné z: doi:10.1111/jedm.12176

desired outcome of testing what should be tested [7]. The same applies to true – false statements, as there is 50% chance to get the response right, which can distort the test results achieved by the candidates[8]. The problem of another fixed-response technique, matching, lies in the fact that possible score might be influenced by the elimination process; nevertheless, this can be to a certain extent restricted by providing an uneven number of items and definitions [9].

On the other hand, constructed response techniques, such as cloze (filling in the gaps/blanks in a passage), short answer questions, or constructed responses, can be used to elicit a more genuine sample of candidate's language. The technique of filling in the gaps has advantages of the short-answer items, but the greater control over likely response(s) does not require significant productive skills. It also calls for a "carefully constructed key on which the scorers can rely completely"[10].

Constructed-response formats are more complex and more demanding to rate; however, they do not constrain the test candidates to the pre-set answers/formulations, allow more accurate picture of candidate's ability, and significantly reduce the danger of guessing the right answer. On the other hand, creating response might take longer resulting in reduction of possible number of items (and "fresh starts"), and candidates have to produce language (not required for receptive skills testing). Also scoring the constructed response tests may be more demanding, it can be unreliable if judgement is required, and it may take longer, and therefore it is less economical[11]. Most of the above-mentioned disadvantages can be overcome by stating the item in such a way that there is only one possible answer and by creating so-called short-answer items. However, creating such a test can be rather difficult and time-consuming.

Each of the above-mentioned techniques (both fixed and constructed responses) has their advantages and limitations; however, they are widely used in language assessment. Although each country surveyed uses different test techniques for testing reading and listening comprehension, all techniques they use are in accordance with contemporary testing theory. In spite of the fact that each country has chosen different techniques to measure the same standard, it can be assumed that the results they achieve are compatible among the countries and they measure the same construct.

---

7 GIERL, M. J., O. BULUT, Q. GUO a X. ZHANG. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*. 2017, **87**(6), 1082–1116. Dostupné z: doi:10.3102/0034654317726529

8 HALADYNA, T.M. *Developing and Validating Multiple-choice Test Items*. 3rd ed. New York: Routledge, 2004. ISBN 9780805846614.

9 HUGHES, Arthur. *Testing for Language Teachers*. 2nd ed. Cambridge: Cambridge University Press, 2013. ISBN 9783125338760; CARR, N. T. *Designing and Analysing Tests*. Oxford: Oxford University Press, 2011. ISBN 9780194422970.

10 HUGHES, Arthur. *Testing for Language Teachers*. 2nd ed. Cambridge: Cambridge University Press, 2013. ISBN 9783125338760

11 Ibid

2.2 Testing productive skills

Testing speaking

The countries of the survey approach assessing speaking skills by using different techniques: role-play, picture description, information gathering task discussion, debate, interview, summary, expressing opinion, hypothesizing. In most countries, presentation (in the form of test taker's monologue) is included in the speaking part of the exam i.a.w. NATO STANAG 6001 of all levels. In some of them, it is an integral part of the introduction or interview, in others it is one of the individual tasks. It allows candidates to produce extended discourse on a given topic, which makes it easier to assess language competence on higher than the sentence level. It also shows the ability to produce logically-structured and coherent speech [12]. Generally speaking, even though the names of the tasks look different, it can be assumed that the ways how the oral examination is conducted are quite similar[13].

Interview can be considered the most commonly used speaking test task (used in all countries). In order to make comparison between test taker's performance and the abilities described by the descriptors, questions used in an interview can be standardized and therefore result in more reliable scoring than when other task types are used[14]. Because of the nature of the discourse that the interview produces, it can be used primarily to test language competence rather than other competencies or knowledge, which is advantageous especially on lower language proficiency levels[15]. However, on higher language competence levels it can be complemented or alternated with presentation technique (see below), which could help to elicit extended discourse samples required on these levels.

Presentation allows the candidate to show his/her ability to produce extended discourse on a certain topic. The danger of candidate's diversion from the set topic to the presentation prepared in advance can be avoided by trained interlocutors.

Picture prompts are another commonly used technique (e.g. Cambridge Assessment) which provides necessary topical support for the test taker. However, there is a danger of test taker not recognizing the situation/context given by a picture – it is important to avoid "culturally alien images" [16]. When selecting suitable picture prompts, it is also important to take into account the expected level of test taker's knowledge as pictures requiring specific language description might prove too difficult for lower levels.

---

12 FULCHER, G. *Testing Second Language Speaking*. London: Routledge, 2003. ISBN 9781315837376; GALACZI, E.D., A. FFRENCH, C. HUBBARD a A. GREEN. Developing assessment scales for large-scale speaking tests: a multiple-method approach. *Assessment in Education: Principles, Policy & Practice*. 2011, **18**(3), 217-237. Dostupné z: doi:10.1080/0969594X.2011.574605; KITAO, S. K. a K. KITAO. Testing Speaking [online]. ERIC, 1996 [cit. 2022-07-22]. Dostupné z: https://eric.ed.gov/?id=ED398260

13 KITAO, S. K. a K. KITAO. Testing Speaking [online]. ERIC, 1996 [cit. 2022-07-22]. Dostupné z: https://eric.ed.gov/?id=ED398260

14 FULCHER, G. *Testing Second Language Speaking*. London: Routledge, 2003. ISBN 9781315837376.

15 Ibid

16 Ibid, p. 75

Picture prompts providing candidates with topical support, and serving as a starting point for presentation and discussion/debate can be replaced by cards with printed situations or with short articles[17]. However, the disadvantage of using the technique is that it requires reading comprehension skills, skills usually tested in a different part of the examination.

Discussion or debate is a widely used speaking assessment technique, as it is an excellent task type for assessing interaction skills, strategic competence, textual and pragmatic knowledge as well as sociolinguistic skills[18]. As the descriptors of level 3 language proficiency exam require a candidate to use language on abstract level, to state and to defend their opinion, as well as to analyze information and create hypotheses, discussion/debate represents a tool that provides examiners with ample space for eliciting the sample needed.

Pairing test takers is common practice in many renowned examination systems (e.g. Cambridge exams). It allows to overcome unequal positions of test taker and interlocutor and its resulting sociolinguistic consequences. Both candidates are on the same level/of the same status, therefore it is easier to oppose their respective opinions or "fight" them. In case of pairing the candidates, it is necessary to allot the same amount of time to each candidate and time strictly their performance in order to prevent one taking "the floor" on behalf of the other [19].

In all countries there are two interlocutors and the candidate's performance is recorded in order to allow the fairest scoring possible: when both scorers are not in full agreement, they can listen again to the performance and there is also a possibility of the third and final opinion of another scorer. This scoring format is in congruence with the expert opinions requesting several independent scorers to assess the performance[20].

It can be concluded that all techniques of testing speaking skills used in particular countries have its justification in specialist literature and can be considered a solid base to produce valid and reliable results.

Testing writing

Writing, as one of the productive skills, is most often assessed by candidate's creating/composing their own text on a given topic, e.g. letter or essay writing, or by controlled writing, using such tasks as e.g. gap-filling, form completion, making corrections [21]. For high-stakes tests it is necessary to find tasks that elicit valid samples of writing, samples

17 Ibid

18 FULCHER, G. *Testing Second Language Speaking*. London: Routledge, 2003. ISBN 9781315837376.

19 Ibid; GALACZI, E.D., A. FFRENCH, C. HUBBARD a A. GREEN. Developing assessment scales for large-scale speaking tests: a multiple-method approach. *Assessment in Education: Principles, Policy & Practice*. 2011, **18**(3), 217-237. Dostupné z: doi:10.1080/0969594X.2011.574605

20 CARR, N. T. *Designing and Analysing Tests*. Oxford: Oxford University Press, 2011. ISBN 9780194422970; HUGHES, Arthur. *Testing for Language Teachers*. 2nd ed. Cambridge: Cambridge University Press, 2013. ISBN 9783125338760; FULCHER, G. *Testing Second Language Speaking*. London: Routledge, 2003. ISBN 9781315837376.

21 KITAO, S. K. a K. KITAO. Testing Speaking [online]. ERIC, 1996 [cit. 2022-07-22]. Dostupné z: https://eric.ed.gov/?id=ED398260

that "truly represent the student's ability" [22]. IELTS mentions description, report, discussion, argument, and opinion as most common text types used in their academic tests, and a letter and an essay on a given topic in their General Test (GT).

As far as the task types in writing part of the examination are concerned, they differ in the countries under survey; also, there are differences based on the level of language proficiency which is tested, what is more, time limits differ according to the country and level.

Kitao[23] identifies two major problems in writing assessment, i.e. the objectivity of evaluation and creating a rating scale which will make "grading as objective as possible". In all countries, the writing is assessed by two testers independently, then they have to consolidate their assessment and agree on one assessment result. In case of disagreement third opinion is required.

As mentioned above, time limits allotted to the tests vary in each skill and each country. Bachman and Palmer [24] distinguish between speeded and power tests. While in speeded tests not all test takers are expected to take all test questions, in power tests enough time should be allotted to allow every test taker to attempt every item. As the tests considered are high-stakes proficiency tests, it is clear that all countries should aim at creating power tests. It is reflected both in test design (number of items/tasks) and in the time allotment.

From all that has already been mentioned it is clear that in most cases (with the exception of one country listening and reading comprehension tests) the countries use multiple/mixed assessment methods that should minimize the bias and technique effect which could undermine the objectivity of the assessment tool and could endanger fairness of the assessment itself [25]. The fairness of the assessment should not be reduced to validity and reliability[26], but broader social and educational context of the assessment and its impact should be considered as well, which is especially true about high-stakes tests (such as tests i.a.w. STANAG 6001).

---

22 Testing writing. In: HUGHES, Arthur. *Testing for Language Teachers* [online]. Cambridge University Press, 2010, 2002-12-12, s. 83-112 [cit. 2022-07-20]. ISBN 9780521484954. Dostupné z: doi:10.1017/CBO9780511732980.010

23 KITAO, S. K. a K. KITAO. Testing Speaking [online]. ERIC, 1996 [cit. 2022-07-22]. Dostupné z: https://eric.ed.gov/?id=ED398260

24 BACHMAN, Lyle F. a Adrian S. PALMER. *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press, 1996. Oxford applied linguistics. ISBN 9780194371483.

25 SHOHAMY, Elana. Fairness in language testing. In: KUNNAN, Anthony John. Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida. [online]. 1. Cambridge: Cambridge University Press, 2000, s. 15-30 [cit. 2022-07-23]. ISBN 978-0521658744; HAMP-LYONS, L. Fairnesses in language testing. In: KUNNAN, A. J. *Fairness and validation in language assessment : selected papers from the 19th Language Testing Research Colloquium*. Cambridge, UK, New York, USA: Cambridge University Press, 2000, s. 30-35. ISBN 9780521658744.

26 MILANOVIC, Michael, ed. Series Editor's Note. In: KUNNAN, Anthony John. Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida. [online]. 1. Cambridge: Cambridge University Press, 2000, vii - viii [cit. 2022-07-23]. ISBN 978-0521658744.

CONCLUSION

The objective of this study was to compare, analyze, and interpret the ways of implementing the same standard into testing processes in Visegrad countries, all of them NATO countries, joining the NATO in the same year, and having similar background in terms of military systems and language training.

The study has revealed a lot of similarities, as well as differences, in the ways how the same standard is being transformed into different language testing systems. The areas in which we have found the biggest similarities are as follows: moderations sessions, pretesting, tester training, requirements for testers' qualifications, use of MCQs in testing receptive skills, use of role-plays and interviews in testing speaking, use of the same writing tasks, the same approach to productive skills assessment (always two raters, if not in agreement, the third rater makes a decision).

On the other hand, a broad range of differences has been revealed, too. To mention the most relevant ones, they are as follows: different testing techniques in measuring both receptive and speaking skills; different number of tasks/ items, different time limits for subtests of different skills and levels; differences in the size of testing teams (6-12) and test takers the number of test takers per year (from approx. 400 – to approx. 4000), and some other.

The reasons for the differences revealed might lie on the macro level, i.e. in the ways how the whole language training is perceived and approached by the General Staffs and/or the Ministries of Defense of respective countries. On the micro level, the reasons behind the differences are in different approaches of the national testing boards, the number of the testing team members, the number of test takers, and many other.

To contribute to better understanding and appropriate application of the standards, it would be beneficial to include more countries in the research, which would enable to create a more precise picture illustrating the same standard use in the NATO countries. This could contribute to better understanding of the assessment process as well as to better assessment co-ordination.

One of the limitations of this study was that although it compared the current techniques, processes, administration, and tester training in the respective countries, it did not compare the actual results of the tests. Further step in the research should therefore be to administer the tests from the chosen countries to a sample of testees, in order to analyze whether different testing processes and methods produce the same results in terms of assessing the level of language proficiency. Conducting the study focused on actual assessment results could shed more light on the studied area.

Therefore, it might be desirable to promote even deeper long-term cooperation including joint test creation and joint testing teams. However, such a decision resulting in far-reaching impact should be taken on governmental, not language expert level.

Authors: ***PhDr. Ivana Mrozková, Ph.D. (1966),*** *graduated from Palacký University, Olomouc, Czech Republic (1989). She has worked as an educator, leader, and scholar in the field of language education, leadership, and leadership communication in the CR and the US. Currently, she works as a head of the Department of Methodology, Language Center, University of Defence, Czech Republic. Her areas of interest range from leadership and leadership communication to language education to language proficiency assessment.*

Mária Šikolová, PhD., *born 1959. She has graduated from the Faculty of Arts of Comenius University in Bratislava, Slovakia (1982) and she defended her PhD thesis at Pedagogical Faculty of Charles University, Prague (1997). She has worked at different management positions in the area of language teaching, testing and methodology, which has been reflected in her publications. She has attended, as well as conducted, numerous workshops and seminars in these areas. Currently she works at the position of a methodologist at the Language Centre of the University of Defence.*

How to cite: MROZKOVÁ, Ivana and Mária ŠIKOLOVÁ. One Standard, Different Approaches: Language Assessment in the Military Context. A Visegrad countries analysis. *Vojenské rozhledy*. 2023, 32 (1), 118-134. ISSN 1210-3292 (print), 2336-2995 (online). Available at: www.vojenskerozhledy.cz