Alternative interpretation of bi-level language proficiency test scores for listening and reading comprehension

Pavel Svoboda, Lenka Marková

Abstract: The aim of this study was to take an alternative look at the language assessment of the listening and reading comprehension with the intention to provide language teachers at the authors' institution with more valid and reliable feedback on their instruction. To achieve this, an alternative approach to the interpretation of test scores obtained from bi-level language proficiency tests assessing the listening and reading comprehension at the authors' institution was analysed. This alternative approach was proposed by the authors for comparison with the currently used approach to test score interpretation at their institution. The intention of the proposed alternative approach was to increase the validity of the proficiency levels obtained based on listening and reading comprehension test scores. In this alternative approach, the authors attempted to eliminate the guessing factor employed by test takers which may negatively affect valid assessment of their listening and reading comprehension. To compare the alternative and the current approaches to the interpretation of test scores, proficiency levels of receptive skills (listening and reading) and productive skills (speaking and writing) were correlated with the assumption that the test score interpretation approach showing higher correlation is more valid. The rationale for this correlation was the fact that the authors considered proficiency levels of productive skills valid and reliable due to the manner in which they are tested at their institution and due to the fact that the proficiency levels of productive skills cannot be distorted by test takers' guessing.

Key words: bi-level language proficiency test; guessing factor; language assessment; test score interpretation; validation

1 Introduction

To provide high-quality and effective language instruction, teachers need valid and reliable feedback on their instruction. Language assessment can provide such valuable feedback, on the condition that it is carried out in a standardised and verified way. The problem of language assessment, however, is that it can only measure the visible output of language knowledge typically within a particular, limited time period and testing occasion, which are constraints that could result in the distortion of the test taker's valid assessment of proficiency. That is why maximum focus should be devoted to minimizing this distortion by seeking out and identifying factors that may negatively affect validity and reliability in the language assessment process.

This is the goal we worked towards in the analysis presented in this paper, since we are professionally involved in the field of language testing and responsible

for language test development, administration, and assessment at our Language Centre, where the language proficiency of Czech military personnel is assessed. In particular, the purpose of this study was to analyse an alternative approach to the interpretation of test scores obtained in bi-level language proficiency tests that assess listening and reading comprehension at our institution. We proposed this alternative approach for comparison with the currently used approach to test score interpretation with the intention of increasing the validity of obtained proficiency levels by eliminating the guessing factor employed by test takers. This guessing factor may negatively affect the valid assessment of listening and reading comprehension, leading thus to a distorted picture of true language proficiency.

Consequently, these proficiency levels with increased validity may provide more precise information for language teachers at our institution concerning the current language proficiency of their students. As a result, the teachers could tailor their language instruction more effectively and focus their efforts on the realistically problematic language skills of their students, and our institution could benefit as military graduates would possess truly valid and high-quality language knowledge required of them in their future military careers and beyond.

The analysis described in this paper could be considered innovative in the sense that similar analyses have not been carried out on this type of language test in a military context. We believe, however, that our findings might be also applicable to other language education contexts and similar types of language tests.

2 Theoretical background

The language test analysed in this study conforms to NATO Standardization Agreement 6001 (STANAG 6001), which responds to NATO's interoperability requirements and aims to "be used as the common standard (construct) for language curriculum and test development, for recording and reporting Standardised Language Profiles (SLPs)" (BILC, 2014, p. 4) in a military context. This language test is a proficiency test. Davis et al. (1990) describe a proficiency test as one measuring "how much of a language someone has learned" (p. 154). Compared to an achievement test, it "is not based on a particular course of instruction" (p. 154). Hughes (2013) points out that this type of test is rather "based on a specification of what candidates have to be able to do in the language in order to be considered proficient" (p. 11).

STANAG 6001 provides detailed descriptions or definitions of five proficiency levels of the commonly recognised language proficiency skills of listening, speaking, reading, and writing (BILC, 2016). The levels of language proficiency are ordered from the lowest level ("Survival"), which practically means minimal knowledge of the language, through "Functional", "Professional" and "Expert" levels, up to the

highest level ("Highly articulate native"), where the language use is equivalent to that of a highly articulate, well-educated native speaker. For the purposes of this paper, only the three lowest proficiency levels were considered since these are the proficiency levels tested at our institution. An overview of these proficiency levels showing simplified level descriptors for all language proficiency skills is provided in the following table.

Tab. 1: *An overview of language proficiency levels (Source: BILC, 2019)*

Level	Description
1 Survival	○ Can understand/produce • simple, routine questions and answers • short phrases within familiar areas to meet immediate personal needs ○ Can participate in simple, short conversations and email exchanges ○ Misunderstandings are frequent
2 Functional	○ Can understand/produce • language for everyday and routine work-related matters • factual accounts of events and activities in present, past and future time • detailed descriptions of people and places • straightforward instructions and directions ○ Uses the language well enough to be generally understood ○ May sound foreign, which sometimes interferes with communication
3 Professional	○ Can understand/produce • formal and informal language for most social and professional situations, e.g. business meetings, conferences, reports on complex issues • well-structured language relating to abstract topics and hypotheses, including technical discussions in his/her field of specialization • detailed arguments for and against different opinions • language to convey implicit information, inferences, and emotional overtones ○ Repetition is rarely requested, has a natural flow, without searching for words ○ Is easily understood by native speakers.

Testing such a wide range of language proficiency levels presents a challenge; therefore, multi-level language proficiency tests are generally developed. According to Abbosov (2023), assessing language skills at multiple levels of proficiency by a single test "leads to more accurate evaluations of language proficiency, which can help teachers and administrators make more informed decisions about placement and instruction" (p. 80). Abbosov also stresses the fact that multi-level testing can lead to improved instruction and curriculum development because by providing a nuanced understanding of language proficiency, teachers and administrators can identify areas of strengths and weaknesses in language instruction. For practicality reasons, however, bi-level proficiency tests are often used instead of multi-level proficiency tests, which are difficult to develop and time-consuming to administer. In bi-level tests, only two levels of language proficiency are tested at the same time as a kind of practical compromise. Unfortunately, both multi-

level and bi-level proficiency tests present challenges in terms of test design and interpretation. Abbosov claims that these tests "must be designed to accurately measure proficiency at each level, while also allowing for comparisons across levels" (p. 80). 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" (p. 81). These challenges are particularly present in the case of proficiency tests assessing listening and reading comprehension skills as the required levels of language proficiency have to be transformed into adequate and valid cut scores.

In this connection, Clifford (2016) notes that listening and reading comprehension skills have traditionally been associated with norm-referenced (NR) testing. NR testing compares test takers' performance within a group, and the test items selected for this type of test primarily distinguish between test takers of varying language abilities. Clifford, however, argues that the criterion-referenced (CR) testing approach is more suitable for assessing listening and reading proficiency. Within CR testing, test takers are not compared with each other, but they are compared against a set of clearly stated expectations or criteria. As a result, all test takers, some test takers, or even no test takers may meet the stated expectations and thus pass or fail a CR test. Clifford specifies that "success requires a sustained, spontaneous ability to perform the communicative tasks that are specified, in the contexts that are described, and with the degree of accuracy that is expected" (p. 225). According to Clifford, a CR test should assess the test taker's language ability at each criterion level separately and then identify the highest level of sustained ability. That is why level-by-level results should not be totalled or averaged across the levels covered by the multi-level test. Clifford claims that the term 'sustained ability' requires an operational definition, and he understands this as "a response pattern that shows sustained or consistently accurate performance at a given level" (p. 230). The sustained performance levels can be established based on expert judgements or with sophisticated statistical procedures; according to Clifford, sustained ability is typically set in the 70 to 80% range, or more specifically, the success criterion is ideally 75%.

The applicability and suitability of the CR testing approach for assessing the receptive skills of listening and reading comprehension have been proved by studies of Clifford and Cox (2013) and Cox and Clifford (2014). In their study, Clifford and Cox (2013) applied a multi-stage, CR approach to validate the ability scales describing the reading skill and related reading proficiency guidelines. The multi-stage approach consisted in the fact that each ability scale was treated and assessed at a separate stage, which is a prerequisite of CR language testing. In their subsequent study, Cox and Clifford (2014) applied the same approach to validate ability scales describing the listening skill and related listening proficiency guidelines. This validation is required for the assessment of both the listening

and reading skills due to the need for the transformation and operationalization of the ability scales into individual test items of difficulty appropriate for each assessed level. Cox and Clifford point out that a multi-level test of a traditional NR design produces a single score for each test taker, and various procedures are used to determine how to convert those scores into proficiency levels. Cox and Clifford consider this process problematic because if a test contains items of varying difficulty, setting an appropriate cut score indicating a proficiency level of the test taker may be challenging. Cizek and Bunch (2006) defined the conversion process as the "concrete activity of deriving cut points along a score scale" (p. 14). By applying CR testing and assessing principles, however, many of the rating deficiencies associated with traditional NR test design can be avoided. Cox and Clifford (2014) for example claim that consequently, "it is no longer necessary to estimate how guessing at levels beyond the test takers' actual ability level may have influenced their total scores" (p. 382).

The factor of guessing of language test takers is an interesting phenomenon, which is reflected in our focus on analysing the effect of this factor in the bi-level tests of listening and reading comprehension skills developed at our institution. We also tried to find possible ways of eliminating the negative impact of this factor on valid interpretation of test scores. Our bi-level tests, developed in accordance with the STANAG 6001 standard, are CR tests in their nature because the test taker's performance is compared against proficiency level descriptors specified within this standard. To minimise the random effect of guessing when taking this type of test and to avoid advantaging some test takers, Davis et al. (1999) advise that all test takers should be told not to guess when they do not know the answer, as opposed to NR tests when all test takers should be told to guess where they do not know the answer. The rationale behind this approach is likely Davis's assumption that in NR tests, the test takers' resorting to guessing to an even extent will not significantly affect their mutual grading when compared to each other. On the contrary, in the case of CR tests, test takers' potential guessing may distort their performance in the form of the obtained test scores, and thus, the proficiency levels awarded on the basis of these scores may not be valid. However, this assumption is controversial, as each test taker may resort to guessing to a different extent depending on many factors. That is why it is crucial to understand the nature of the complex phenomenon of guessing in order to be able to achieve its elimination.

Experts treat guessing in various ways; for example, Bachman (1990) considers guessing one of the parameters of test items that can be described by item characteristic curves in Item Response Theory models. According to Bachman, this 'guessing' parameter expresses "the probability that an individual of low ability can answer the item correctly" (p. 204). Alderson (2000) used multiple methods for examining test items, one of which was applying think-aloud techniques to

identify the strategies that test takers used. As a result, he found that test takers reported guessing with difficult items being one of their test-taking strategies. As for a practical definition of guessing, for example Davis et al. (1999) define it as "the apparently random selection of an answer (which may involve absolute randomness or may first involve elimination of those alternatives known to be wrong)" (p. 70). As a result, guessing introduces a factor of randomness into test scores which then lowers reliability and validity of these scores. Davis et al. also claim that guessing may cause overestimation of the test takers' ability if they guess correctly.

The problem of guessing is typically connected with selected-response (or forced-choice) items wherein the test takers are presented with a small, finite set of options from which they must choose (e.g. true/false or multiple-choice items). For such items, the probability of accidental success can be calculated and correction for guessing can be employed. Some experts, however, criticise this procedure as being too simplistic and consider it "problematic in itself as it assumes that all candidates are affected by guessing to the same degree" (Davis et al., 1999, p. 35). Bachman and Palmer (1996) point out that there are several factors that must be considered when employing correction for guessing, one of them being the fact that "the tendency to guess is affected by a number of different personality characteristics and varies greatly from one individual test taker to the next" (p. 204). Other factors are, for example, the test taker's level of ability and the nature of the test task itself. Bachman and Palmer conclude that correction for guessing is virtually never useful in language tests, and they recommend "to include, in the test design, provisions for eliminating or reducing the potential causes of random guessing, rather than correcting for it after the fact" (p. 205). According to them, one of these provisions is to match the difficulty of items with ability levels of test takers. To minimise guessing in tests, Davis et al. (1999) claim that forced-choice items should be avoided, and Hughes (2013) suggests that if multiple-choice items are to be used in a test, "every effort should be made to have at least four options (in order to reduce the effect of guessing)" (p. 77). Relating to this, Carr (2011) argues that for a four-option multiple-choice item, the likelihood of test takers answering correctly is actually lower than the expected 25% because truly random guessing is rare.

Consequently, forced-choice items are acceptable in NR tests because they do not significantly affect the validity of test scores; however, in CR tests, this type of test item should be avoided. Instead, constructed-response items should be favoured because these require test takers to formulate their own responses and thus preclude any random effect of guessing. Nevertheless, considerations of practicality usually force test developers to resort to forced-choice items even in CR tests as the scoring of these items is very easy and time efficient. With the spread of electronic testing, it is often the only suitable item type that the software can

reliably cope with. As a result, the test developers must look for other possible ways to eliminate the undesirable effect of guessing in their tests.

We tried to analyse this effect of guessing in our tests, which we develop in accordance with STANAG 6001. This document only provides a language proficiency standard for NATO countries while preserving each nation's right to maintain its own internal proficiency standards and interpretation of this standard (BILC, 2016), meaning that each NATO country can develop its own test versions in accordance with STANAG 6001. In the Czech Republic, the institution responsible for the development of these tests is the Language Centre of the University of Defence located in Brno, and the purpose of these tests is to assess the English language proficiency of NATO military and civilian personnel who are non-native users of English. It means that the test takers include students of the University of Defence, where we are professionally active as full-time test developers and administrators within the Testing Department of the Language Centre. This is the reason for our interest in providing the University language teachers with valid and reliable feedback on their students' test results.

Based on our test specifications (University of Defence Language Centre, 2019), the test is a criterion-referenced proficiency test that measures proficiency in accordance with STANAG 6001 regardless of the specific circumstances in which language ability was acquired. The test offers a sufficiently large and varied range of tasks to provide an adequate language sample demonstrating what test takers are able to do in the target language, and their performance is measured against the STANAG 6001 descriptors. The test is administered almost every workday throughout the year, and the annual number of test takers typically reaches 4,000–4,500. The levels assessed at our institution are level 1 ("Survival"), level 2 ("Functional"), and level 3 ("Professional"). The test assesses language proficiency in two receptive skills, listening and reading comprehension, and two productive skills, speaking and writing. The analysis in our paper focuses on bi-level tests of listening and reading comprehension that are used for levels 1–2 and 2–3. For practicality reasons, all tests consist of multiple-choice items with one correct answer and three distractors. The format of the tests is shown in the following table.

Tab. 2: *Format of the listening and reading bi-level tests*

Listening/Reading comprehension	Number of test items (30 items in each test)
Level 1–2 tests	15 level 1 items 15 level 2 items
Level 2–3 tests	15 level 2 items 15 level 3 items

The listening items are usually semi-authentic recordings for lower levels and authentic, unedited recordings for the higher levels. The test takers hear each recording twice, with adequate pauses provided for selecting an answer. The tasks that the test takers are required to do in the listening and reading comprehension tests are detailed in the following table.

Tab. 3: *Test tasks in listening and reading comprehension tests (Source: University of Defence Language Centre, 2019)*

Proficiency level	Test tasks
Level 1	• Understand the main idea • Identify persons, places, things and time related to the main idea
Level 2	• Understand the main idea • Understand specific information/details
Level 3	• Understand the main idea • Understand specific information/details • Make inferences and deductions • Understand the author's intention, attitude or tone • Understand hypothesis, analysis, argumentation, and various forms of elaboration

As for the duration of the bi-level tests, the listening tests are assigned time depending on the length of the individual listening items. For the listening level 1–2 tests, it is typically around 45 minutes, while for the listening level 2–3 tests it is around 60 minutes. In the case of reading tests, the time allotment is 35 minutes for the level 1–2 tests and 65 minutes for the level 2–3 tests. The tests are computer-based, and they are administered through a system called ETIS (Electronic Testing and Information System), which is especially tailored to the needs of our institution. This system, for example, enables the mixing of the item options for each test taker, preventing cheating in this way, and it also allows for the individual pace of each test taker, resulting in different times needed by the test takers for completing the test. When the test is completed by all test takers, the test administrators immediately know test scores of the individual test takers, and they can assign the appropriate proficiency levels based on these scores, as it is shown in the following table.

The cut scores are based on earlier statistical analyses, as well as on the psychometric characteristics of the test items (the Angoff method, Thompson, 2024) previously obtained at our institution. It can be seen in the table that the cut

Tab. 4: *Proficiency levels based on bi-level test scores*

Level 1–2 Listening/Reading comprehension tests		Level 2–3 Listening/Reading comprehension tests	
Test scores	**Proficiency level**	**Test scores**	**Proficiency level**
0–12	0	0–12	N
13–17	1	13–17	2
18–21	1+	18–20	2+
22–30	2	21–30	3

scores for the highest proficiency levels of the bi-level tests are different for level 1–2 and 2–3 tests, which reflects the fact that level 2–3 tests are quite demanding and that the lower cut score is justifiable for them. The "+" proficiency levels are derived from STANAG 6001, and they account for a proficiency level that is nearly at the level of the next base level while not entirely or consistently meeting all of the criteria for that higher level. The "N" proficiency level in the case of level 2–3 tests stands for "not assessed" because these bi-level tests do not contain items of the appropriate difficulty level upon which it could be possible to validly decide what is the proficiency level of the test takers not achieving at least level 2 proficiency.

As for the productive skills of speaking and writing assessed by our bi-level tests, their format is described in the following table.

During the speaking test, only one test taker is tested at a time under standardised conditions, and two testers always participate in evaluating the test taker's performance in the speaking and the writing tests. The performance is rated holistically using the STANAG 6001 descriptors, meaning that the testers look for general patterns of strengths and weaknesses, and they award the appropriate proficiency level. Consequently, test scores are of no importance for this rating. The testers rate the performance independently and in case of a discrepancy, a senior tester is asked to rate the performance. All testers are well-trained in rating, and they participate in regular calibration sessions where their ratings are compared to ensure validity and reliability.

As for validity, we consider it to be crucial in language assessment and its assurance should be the main goal for test developers. This is the reason why validity is widely discussed by many testing experts. The concept of validity was for example analysed in an innovative way by Borsboom et al. (2004), and Xi (2010) stresses the importance of fairness as an aspect of validity. In general, the notion of validity has shifted in the language testing handbooks from the classical definition of Davis et al. (1999) who consider a test to be valid when it "really measures what it claims to measure" (p. 221) to the broader definition of validity when the test and also its real-life context are considered. In connection with this, Green (2014)

Tab. 5: *Format of the speaking and writing bi-level tests (Source: University of Defence Language Centre, 2023)*

	Speaking	Writing
Level 1–2 tests	• Introduction: A conversation on concrete topics • Role-play: Simulating an everyday situation with or without complication. The role-play is an interaction between one examiner and the test taker • Information Gathering Task (IGT): Asking one of the examiners for information and reporting the information to the other examiner (10–15 minutes)	• 2 prompts • Task 1 (Level 1): basic correspondence, such as a short note, post card, short personal letter, phone message, invitation (70 words) • Task 2 (Level 2): simple personal, social and routine workplace correspondence, such as a memorandum, brief report, personal letter, complaint (150 words) (40 minutes)
Level 2–3 tests	• Introduction: A conversation on concrete topics in everyday social and routine workplace situations • Interview: A conversation on social and professional topics on both concrete and abstract levels • Debate: Presenting and supporting one's viewpoint on a given topic; the test taker is expected to clearly explain his/her views on both concrete and abstract levels, and he/she should also respond to counter-arguments of the examiner. (20–25 minutes)	• 2 prompts • Task 1 (Level 2): work-related or personal correspondence (150 words) • Task 2 (Level 3): essay-like composition, a choice of two topics – one is military based, the other of general interest (300 words) (90 minutes)

points out that a test cannot be valid on its own, and validity is not a quality of a test, but "it is a quality of the interpretations that users make of assessment results: an assessment can only be considered valid for a certain purpose" (p. 75). Consequently, Green also claims that "validity is better seen as a matter of degree" (p. 76) because inferences made from test results may be more valid for one type of stakeholder decision than for another, and he defines validation as "the process of collecting evidence of the validity of inferences based on assessment results" (p. 76).

This approach reflects the developments in argument-based approaches to validation in language assessment where authors such as Kane (2002, 2006, 2013) deal with the links between test takers' performance, test score interpretations, and test score uses. In Kane's interpretative argument, validation of the interpretations and uses of test scores means the evaluation of the plausibility of the claims based on the scores. Consequently, he considers test-score interpretations and uses to be

valid when they are clearly stated and supported by appropriate evidence. In her assessment validation, Ryen (2002) also includes the perspectives of stakeholder groups, such as teachers or parents, claiming that this can improve the validity of high-stakes assessment interpretations and uses. She argues that the views of these stakeholders can contribute to identifying the strengths and weaknesses of the intended assessment interpretations and uses. In his validity theory, Cizek (2012) points out the incompatibility of incorporating score meaning and score use into a single concept, and he proposes a framework that both accommodates and differentiates validation of test score inferences and justification of test use.

Bachman (2002, 2005) and Bachman and Palmer (2010) elaborate the concept of validity in which they emphasise the importance of the use of language assessments. They consider the primary use of an assessment to be the gathering of information to aid in making decisions that will lead to beneficial consequences for the stakeholders. On the basis of this premise, they have developed a conceptual framework called the Assessment Use Argument (AUA). This framework describes assessment use as a series of inferences, or links, from a test taker's performance on the assessment tasks to the intended decisions and consequences of the assessment. At the same time, it describes the process of assessment justification- that the test developers should follow to justify their decisions in designing and developing an assessment. The AUA validation framework can thus guide both the test development and the test use validation process, as shown in the following figure.

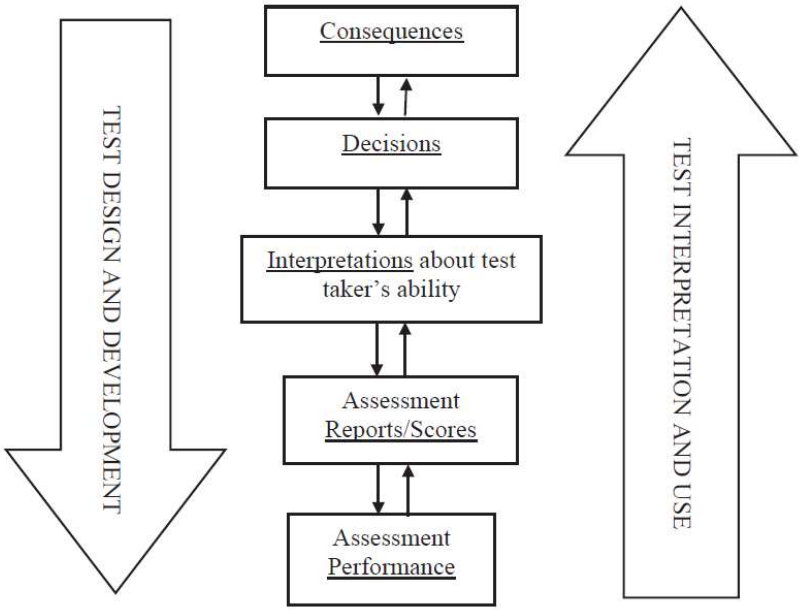

Fig. 1: *Inferential links from consequences to assessment performance (Source: Bachman and Palmer, 2010, Fig. 5.1, p. 91)*

This validation framework provided us with a theoretical concept, upon which we based our analysis wherein we primarily focused on the central part of the framework concerning validation of the interpretations about the test taker's language ability.

3 Methodology

3.1 Research design

To achieve the aim of validating our proposed alternative approach to the interpretation of bi-level language proficiency test scores for listening and reading comprehension, we elected, in accordance with Green (2014), Kane (2002, 2006, 2013), and Bachman and Palmer (2010), to gather and evaluate the evidence for the validation of our alternative approach to test score interpretation by investigating criterion-related validity. According to Davis et al. (1999), this type of validity of a new test is analysed statistically (using correlation) in terms of the closeness of the new test to its criterion, which may be an established and valid test or some other measure within the same domain. Davis et al. characterise this type of criterion-related validity as concurrent validity, wherein the two measures having this validity are considered to function similarly. For us, the valid criterion against which we measured the concurrent validity were the speaking and writing tests administered at our institution as these tests are highly standardised in accordance with STANAG 6001, the proficiency levels in these tests are awarded directly on the basis of the descriptors specified within STANAG 6001, the testers regularly attend calibration sessions, the performance of each test taker is assessed independently by two testers, and the results of these tests cannot be affected by the test takers' guessing. Consequently, we consider all these facts adequate evidence ensuring a high validity of the awarded speaking and writing proficiency levels. The other measure that was compared and correlated with this valid criterion was the approach to the interpretation of bi-level language proficiency test scores for listening and reading comprehension on the basis of which proficiency levels for these receptive skills are awarded.

The rationale for this correlation was our assumption that both the productive and receptive language skills belong to the same language domain where they closely interoperate and affect each other. This assumption is supported by the study of Lakshani (2015), in which he correlated scores for the receptive and productive skills in a placement test obtained by learners of English as a second language. A moderate positive correlation was found, and Lakshani concluded that sound proficiency in the receptive skills paves the way for higher productive skill proficiency in learning English as a second language. A similar analysis was carried out by Yazar and Rejeki (2020) within a military context where they analysed the connection between the receptive and productive skills in a diagnostic pro-

ficiency test assessing army officers' English language competence. The analysis showed a strong correlation between the scores of receptive and productive skills, while listening proved to be the dominant skill that correlated strongly with all other proficiency skills.

However, we are also aware of the fact that in the context of second language (L2) learning, a 'jagged' language profile can sometimes appear that refers to the uneven development of language skills in areas such as speaking, listening, reading, and writing. According to Kostromitina (2024), this can manifest differently depending on various factors, including individual differences among the test takers or language learning and teaching methods. Kostromitina claims that understanding jagged language profiles in L2 acquisition highlights the importance of tailored instruction and testing. From the teaching perspective, she points out that "recognizing learners' strengths and weaknesses across various language domains allows instructors to provide targeted support and interventions to address specific areas of need, ultimately promoting a more effective language learning experience" (p. 1). As for the testing, she claims that "acknowledging the variability in language profiles is crucial for language testing in order to provide a more precise and pleasant testing experience and accommodate test takers' differences in language skills" (p. 1).

Based on the above-mentioned studies, we expected the correlation between the productive and receptive skills to be high, with some possible exceptions caused by test takers having jagged language profiles of certain skills. Having this in mind, we compared two approaches to the test score interpretation of the receptive skills: one currently used at our institution and the alternative one we propose. In accordance with Davis et al. (1999), Green (2014), and Kane (2006), we considered the approach providing higher correlation with the valid criterion as having higher concurrent validity. Consequently, this approach would be more justified for test score interpretation use because the proficiency levels in receptive skills obtained by the test takers (i.e. the University students) would provide their teachers with more valid and reliable feedback on their instruction. Ideally, the decisions made by the teachers would thus be fairer and more tailored, and the consequences would be more beneficial for the stakeholders, as is required in the Assessment Use Argument by Bachman and Palmer (2010).

3.2 Data collection

The first step was to collect the test takers' results of the tests in the receptive skills of listening and reading and the productive skills of speaking and writing administered at our institution. In terms of receptive skills, four different test versions of the language proficiency test were chosen for the analysis: two bi-level test versions testing listening and reading proficiency at levels 1 and 2, and two

bi-level test versions testing these proficiencies at levels 2 and 3. For the purpose of this paper, the tests were assigned the following codes: 1–2T1, 1–2T2, 2–3T3 and 2–3T4. The test takers' results were obtained from the database of the ETIS testing software, on the basis of which the appropriate proficiency levels were assigned (see Table 4). The results analysed in this study were collected for the period from January 2021–March 2022. The sample size for the individual test versions was 282 test takers for 1–2T1, 296 test takers for 1–2T2, 177 test takers for 2–3T3 and 153 test takers for 2–3T4. The difference in sample sizes was caused by the fact that the lower proficiency tests were needed and used more often during the examined period than the higher proficiency ones. The test takers were of various type, but most typically they comprised University of Defence students, career soldiers, and civilian employees working for the military. As for speaking and writing skills, the results for the same test takers in the form of their language proficiency levels were obtained from the Academic Affairs Department at our institution, where records concerning the test takers' achieved proficiency levels in the tests are kept.

3.3 Analysis

In the next phase, it was necessary to find out whether the proposed test score interpretation approach is really more valid than the current interpretation approach. The innovativeness of the former approach was the fact that is was based on the principles of the criterion-referenced testing approach described in the theoretical part of this paper and, in accordance with recommendations and findings of Clifford (2016), Clifford and Cox (2013), and Cox and Clifford (2014), we decided to interpret the score results for listening and reading skills of the bi-level language proficiency test separately for each tested proficiency level. On the contrary, in the current approach, the test taker's proficiency level is awarded on the basis of an overall score totalled for both tested proficiency levels. As the bi-level test is, for practical reasons, composed solely of multiple-choice items, the current interpretation approach does not take into consideration the fact the test takers with lower language proficiency may tend to simply guess the correct answers for items with a difficulty beyond their language abilities and thus distort their true overall score.

Following the recommendation of Bachman and Palmer (1996) that one of the possible provisions how to reduce the potential causes of random guessing is matching the difficulty of items with the ability levels of test takers, we proposed a two-stage score interpretation approach, in which the test taker's performance in the lower proficiency subtest was analysed in the first stage. After that, only if the test taker proved sustained ability at the lower level of proficiency was also his or her performance in the higher proficiency subtest considered and assessed in the second stage. The crucial point was then to decide what sustained ability

really means, how to operationalise it into an appropriate cut score, and how to translate the test scores into valid proficiency levels.

Using the concept of sustained ability described by Clifford (2016), we tried to find the cut score best representing sustained ability and, in compliance with Clifford's recommendation, the ideal success criterion level was set at 75%. To cover a wider range of sustained ability, we also decided to examine the 70% and 80% success criterion levels for comparison. Transformed into appropriate cut scores, this meant 10-, 11-, and 12-point cut scores out of a maximum of 15 points for each proficiency level. Then, on the basis of the proposed two-stage score interpretation approach, proficiency levels were awarded, as the example for the 75% success criterion level shows in the following table. The other two success criterion levels worked analogically.

Tab. 6: *Proficiency levels based on the 75% success criterion level (11-point cut score)*

Level 1–2 Listening/Reading comprehension tests				**Level 2–3 Listening/Reading comprehension tests**			
Level 1 test scores (1st stage)	**Level 2 test scores (2nd stage)**	**Proficiency level**	**Numeric code**	**Level 2 test scores (1st stage)**	**Level 3 test scores (2nd stage)**	**Proficiency level**	**Numeric code**
0–10	ignored	0	0	0–10	ignored	N	0
11–15	0–6	1	1	11–15	0–6	2	1
11–15	7–10	1+	2	11–15	7–10	2+	2
11–15	11–15	2	3	11–15	11–15	3	3

In this way, alternative proficiency levels for the test takers' listening and reading skills were obtained and, consequently, it was necessary to ascertain whether these proficiency levels provided more accurate and valid information about the test takers' language abilities than the proficiency levels awarded on the basis of the current interpretation approach. For this purpose, we correlated the test takers' proficiency levels in receptive skills obtained through both score interpretation approaches with the proficiency levels of the same test takers obtained in productive skills. In compliance with the concept of concurrent validity, we assumed that the score interpretation approach showing a higher correlation with proficiency levels in productive skills is more valid because these proficiency levels can be considered a valid criterion measure as reasoned above, and they are furthermore not distorted by the test takers' guessing.

4 Findings

After the correlation of proficiency levels in receptive and productive skills, a set of correlation coefficients was obtained for each of the four examined tests, as the example for the test 1–2T1 shows in the following table.

Tab. 7: *Correlation coefficients for test 1–2T1*

Test 1–2T1	Speaking	Writing
Listening	0.359	0.319
70% Listening	0.382	0.331
75% Listening	0.383	0.347
80% Listening	0.382	0.337
Reading	0.431	0.329
70% Reading	0.473	0.369
75% Reading	0.410	0.311
80% Reading	0.437	0.405

Each set contains coefficients of correlation between proficiency levels in productive and receptive skills where the proficiency levels in receptive skills were based on the current score interpretation approach (lines starting with "Listening" and "Reading"), and the lines below contain coefficients of correlation between proficiency levels in productive and receptive skills based on the proposed interpretation approach. These lines start with an appropriate success criterion level that was applied to each subtest of the bi-level tests. The highest correlation coefficients for each correlation group were highlighted, and for all four tests, the highest coefficients were always achieved for some of the examined success criterion levels. Conversely, the same did not happen for any of the coefficients based on the current interpretation approach. Quantitatively, the number of cases with the highest coefficients achieved for each success criterion level is shown in the following table.

Tab. 8: *Number of cases of the highest coefficient*

Success criterion level	Number of cases of the highest coefficient
70%	3
75%	7
80%	6

As seen in Table 8, the 75% success criterion level provided the best correlation between the productive and receptive skills and, consequently, it was necessary to find out whether the correlation coefficients obtained with this interpretation approach are statistically significantly higher than those coefficients provided by the current interpretation approach. For this purpose, all 16 pairs of the correlation coefficients for the current interpretation approach and for the 75% success criterion level of the proposed interpretation approach were collected in the following table to be compared and examined. The differences between the coefficients in pairs were also calculated to judge to what extent the proposed interpretation approach provides higher or lower correlation coefficients than the current interpretation approach.

Tab. 9: *Pairs of correlation coefficients for the 75% success criterion level*

	Current interpretation approach	Proposed interpretation approach	Difference	Test statistic z	Probability p (one-tailed)
Test 1–2T1	0.359	0.383	0.024	0.329	0.371
	0.319	0.347	0.028	0.372	0.355
	0.431	0.410	−0.021		
	0.329	0.311	−0.018		
Test 1–2T2	0.394	0.481	0.087	1.304	0.096
	0.277	0.353	0.076	1.022	0.153
	0.417	0.435	0.018	0.266	0.395
	0.337	0.396	0.059	0.825	0.205
Test 2–3T3	0.193	0.252	0.059	0.579	0.281
	0.176	0.258	0.082	0.803	0.211
	0.297	0.387	0.090	0.952	0.171
	0.290	0.424	0.134	1.436	0.075
Test 2–3T4	0.159	0.147	−0.012		
	0.278	0.266	−0.012		
	0.286	0.368	0.082	0.796	0.213
	0.353	0.448	0.095	0.981	0.163

In the next phase, the Fisher r-to-z transformation was applied to calculate the value of z used to assess the significance of the difference between two correlation coefficients of two independent samples. This value was calculated only for those pairs of coefficients where their difference was positive as it would otherwise be meaningless to examine the significance of their difference (highlighted cases in Table 9). For this purpose, an online calculator was used, which also provided the value of probability p showing the probability of rejecting the null hypothesis (i.e. there is no significant difference between the two correlation coefficients in pairs) when the null hypothesis is true. All the calculated values are shown in Table 9 and with the required significance alpha level set at 0.05, it means that all p-values surpassed this significance level.

One additional analysis was carried out in which the proficiency levels for both productive skills and both receptive skills were integrated into two respective proficiency values for each test taker. The obtained values were correlated for each examined test and for the individual success criterion levels to find out whether productive skills in general correlate better with receptive skills when the proposed interpretation approach is used, or when the current interpretation approach is used. Correlation coefficients obtained in this way are shown in the following table.

Tab. 10: *Coefficients of correlation between receptive and productive skills*

	Productive skills			
	1–2T1	**1–2T2**	**2–3T3**	**2–3T4**
Receptive skills	0.469	0.439	0.348	0.368
Receptive skills 70%	0.499	0.467	0.461	0.425
Receptive skills 75%	0.472	0.513	0.475	0.410
Receptive skills 80%	0.510	0.502	0.465	0.415

It is evident from the table that all correlation coefficients obtained by the proposed interpretation approach (70%, 75% and 80% success criterion levels) are higher for all examined tests than the correlation coefficients obtained by the current interpretation approach. The highest value of these coefficients for each test is highlighted in Table 10, showing that for the 70% and 80% success criterion levels, the highest coefficient was achieved once, and for the 75% success criterion level, it was achieved twice. This suggests that the 75% success criterion level provides the best correlation between the receptive and productive skills. To test the statistical significance of the fact that this success criterion level of the proposed interpretation approach provides more valid assessment results than the current interpretation approach, appropriate pairs of correlation coefficients were compared and their difference examined, as shown in the following table.

Tab. 11: *Pairs of correlation coefficients for the 75% success criterion level (receptive and productive skills integrated)*

	Current interpretation approach	**Proposed interpretation approach**	**Difference**	**Test statistic** z	**Probability** p **(one-tailed)**
1–2T1	0.469	0.472	0.003	0.046	0.482
1–2T2	0.439	0.513	0.074	1.160	0.123
2–3T3	0.348	0.475	0.127	1.430	0.076
2–3T4	0.368	0.410	0.042	0.429	0.334

The table shows that for all the examined tests, the difference between the correlation coefficients is positive, but in some cases, the difference is only minimal. Furthermore, the *p*-values were calculated and for the significance alpha level set at 0.05, all of these *p*-values surpassed the required level.

5 Discussion

The results of the statistical analysis suggest the proposed score interpretation approach for the receptive skills provides more valid test taker proficiency levels than the current interpretation approach because the correlation coefficients between the receptive and productive skills are higher overall when the proposed interpretation approach is applied. From the three examined success criterion

levels of this alternative interpretation approach, the 75% success criterion level provides the best correlation coefficients; however, in the case of two examined tests (1–2T1 and 2–3T4 as shown in Table 9), these correlation coefficients were lower for the correlation of some skills than when the current interpretation approach was applied. This may have been caused by a number of factors, some of which may have been the content and difficulty of the particular test, varying language proficiency of the test takers in the particular sample, and the size of the particular sample of test takers.

From the point of view of statistical significance, however, not even for the 75% success criterion level of the proposed interpretation approach were the obtained correlation coefficients sufficiently higher than the correlation coefficients provided by the current interpretation approach because the appropriate p-values were always higher than the significance alpha level set at 0.05. This means that it is not possible to accept the alternative hypothesis that the proposed interpretation approach provides more valid proficiency levels than the current interpretation approach with adequate confidence.

Similar results were achieved when the language proficiency levels for both the receptive and productive skills were integrated into two respective proficiency values for each test taker and these values were correlated. The 75% success criterion level of the proposed interpretation approach again provided the best correlation coefficients (shown in Table 10); however, all p-values calculated for this success criterion level (Table 11) surpassed the significance alpha level set at 0.05. For this reason, it is not possible to confidently accept this success criterion level as a better way of test score interpretation than the current interpretation approach.

6 Conclusion and further research

We proposed and analysed an alternative listening and reading test score interpretation approach applied to bi-level language proficiency tests. This approach was proposed as an alternative to the interpretation approach currently used at our institution and was based on a two-stage interpretation approach. In both of these stages, language performance of the test takers was assessed separately at each proficiency level of the bi-level test, and on the basis of these two test score interpretations, the overall language proficiency of the test takers was determined. The purpose of the proposed interpretation approach was the intention to eliminate the factor of guessing by test takers, which we had supposed must be resorted to when dealing with test items beyond their language abilities, potentially distorting assessment results. This alternative interpretation approach was thus supposed to provide more valid proficiency levels of the test takers.

Besides improving the quality of language assessment at our institution, application of the proposed alternative test score interpretation approach could also result in a more positive testing experience for the test takers themselves. This could be achieved by software adaptation of the ETIS testing system used at our institution that would shorten the test takers' examination length in listening and reading bi-level tests if they do not prove their sustained language ability in the lower proficiency level of the test. In such a case, the system would prevent them from continuing in their assessment performance and in proceeding to the higher proficiency level of the test. In this way, the test takers' anxiety would be eliminated by not being exposed to the part of the test that is beyond their true language ability. This fact could also result in reducing the risk of this part of the test being compromised. This adaptation of the ETIS testing system could also represent a type of precursor of the potential adaptive language testing possibly introduced at our institution in the future.

The analysis described in this paper, however, failed to provide adequate statistical support in favour of the proposed interpretation approach, in spite of this concept providing some parameters supporting its benefits. The result of the analysis may have been caused by the fact the test takers' guessing is not affecting their performance to such an extent as had originally been expected and, consequently, the guessing factor is not significantly destructive for the determination of the test takers' valid language proficiency in their receptive skills. The test takers likely resort to various helping strategies by which they avoid or reduce the necessity to solely guess when dealing with difficult test items.

Although not proved statistically, our proposed interpretation approach shows some positive features, and it indicates that the criterion-referenced testing approach to assessing listening and reading proficiency by multi-level tests (as recommended by Clifford, 2016) is adequate and beneficial. From the point of view of the Assessment Use Argument validation framework (Figure 1), the test design and development of similar multi-level proficiency tests should follow the criterion-referenced testing principles to provide valid and reliable results. As for the test interpretation and use, the obtained test results could prove useful to University of Defence language teachers as a form of fair and valuable feedback on their instruction. On the basis of such feedback, they could take appropriate and better-informed decisions to tailor their language instruction more effectively with beneficial consequences not only for their students, but for the University itself, providing high-quality language instruction and having graduates who would be linguistically prepared for their future military careers.

The potential benefit for language teachers and their students would be the fact that the language assessment would reflect the students' actual language proficiency with higher validity and reliability. On the basis of such assessment, deci-

sions could be taken so that students complying with the language requirements of the University could focus their language learning efforts and time either on English language acquisition at a higher level, or on increasing their proficiency in a particular language skill or another language of their preference. In all these cases, language teachers could thus tailor their language instruction adequately, making it a more efficient and enjoyable experience, both for their students as well as themselves.

Being aware of the limitations of our study, our analysis could be expanded in the future to include larger samples of test takers and a higher number of tests to reduce possible factors distorting the results of the analysis. A more targeted approach could also be taken in further research to focus attention on specific criteria of our proposed test score interpretation approach. Our study could thus be taken as another step in our effort to improve the quality of language assessment that our Language Centre provides for the University of Defence.

References

Abbosov, A. A. (2023). Analysis of multilevel English language proficiency tests. *Bulletin*, *2023*(1), 79–96.

Alderson, J. Ch. (2000). Assessing reading. Cambridge University Press.

Bachman, L. (1990). Fundamental considerations in language testing. Oxford University Press.

Bachman, L., & Palmer, A. (1996). Language testing in practice. Oxford University Press.

Bachman, L. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, *21*(3), 5–18.

Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*, 1–34.

Bachman, L., & Palmer, A. (2010). Language assessment in practice. Oxford University Press.

BILC. (2014). STANAG 6001. (Edition 5). [online]. Dostupné z: https://www.natobilc.org/files/file/6001EFed05.pdf

BILC. (2016). STANAG 6001 language proficiency levels. (Edition A, Version 2). [online]. Dostupné z: https://www.natobilc.org/files/ATrainP-5%20EDA%20V2%20E.pdf

BILC. (2019). Overview of language proficiency levels. (5th ed.). [online]. Dostupné z: https://www.natobilc.org/documents/TrainingResources/STANAG%206001%20Overview%20Feb%202019.pdf

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

Carr, N. (2011). Designing and analyzing language tests. Oxford University Press.

Cizek, G. J., & Bunch, M. (2006). Standard setting: a guide to establishing and evaluating performance standards on tests. SAGE Publications.

Cizek G. J. (2012). Defining and distinguishing validity: interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31–43.

Clifford, R. (2016). A rationale for criterion-referenced proficiency testing. *Foreign Language Annals*, *49*(2), 224–234.

Clifford, R., & Cox, T. (2013). Empirical validation of reading proficiency guidelines. *Foreign Language Annals*, *46*(1), 45–61.

Cox, T., & Clifford, R. (2014). Empirical validation of listening proficiency guidelines. *Foreign Language Annals*, *47*(3), 379–403.

Davis, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). Dictionary of language testing. (1st ed.). Cambridge University Press.

Green, A. (2014). Exploring language assessment and testing. Routledge.

Hughes, A. (2013). Testing for language teachers. Cambridge University Press.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41.

Kane, M. (2006). Validation. In R. L. Brennan (ed.), *Education Measurement* (4th ed.), 17–64, American Council on Education and Praeger Publishers.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*,*50*(1), 1–73.

Kostromitina, M. (2024). Is this normal? Parsing jagged profiles in language assessment. [online]. Dostupné z: https://blog.englishtest.duolingo.com/is-this-normal-parsing-jagged-profiles-in-language-assessment/

Lakshani, J. (2015). Correlational study – The role of receptive language skills in the acquisition of productive skills. [online]. Dostupné z: https://www.researchgate.net/publication/344188030_Correlational_Study_-The_Role_of_Receptive_Language_Skills_in_the_Acquisition_of_Productive_Skills

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice, 21*(1), 7–15.

Thompson, N. (2024). Modified-Angoff Method Study. [online]. Dostupné z: https://assess.com/conduct-a-modified-angoff-study/

University of Defence Language Centre. (2019). Manual for item development. [online]. Dostupné z: https://intranet.unob.cz/rektorat/cjv/_layouts/15/WopiFrame.aspx?sourcedoc=%7B2AE1BD79-2F03-4D7C-8134-DFDB930F1838%7D&file=Manual%20for%20item%20development_2019.doc&action=default&CT=1708895075988&OR=DocLibClassicUI

University of Defence Language Centre. (2023). Test specifications. [online]. Dostupné z: https://intranet.unob.cz/rektorat/cjv/Dokumenty%20OMTJ/Forms/AllItems.aspx?RootFolder=%2Frektorat%2Fcjv%2FDokumenty%20OMTJ%2Fdokumenty%2Ftvorba%5Fmoderace%5Ftestovych%5Fuloh&View=%7BF8744BFC%2D3561%2D4802%2DB7DE%2DB8121E585361%7D

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170.

Yuzar, E., Rejeki, S. (2020). Correlation between productive and receptive language skills: An examination on ADFELPS test scores. *SALEE: Study of Applied Linguistics and English Education*, *1*(2), 99–113.

Authors

Mgr. Pavel Svoboda, e-mail: pavel.svoboda@unob.cz, Oddělení testování, Centrum jazykového vzdělávání, Univerzita obrany, Kounicova 65, 662 10 Brno, Czech Republic
Pavel Svoboda is an English language tester and administrator at the Testing Department of the University of Defence Language Centre in Brno, Czech Republic. He has extensive experience both in English language testing and instruction at the Language Centre. His primary interest is in the development of English proficiency tests in accordance with STANAG 6001, entrance tests for the University of Defence, and placement tests for the needs of the Czech military. He also focuses on the statistical analysis of English proficiency and placement tests.

Mgr. Lenka Marková, B.A., e-mail: lenka.markova@unob.cz, Oddělení testování, Centrum jazykového vzdělávání, Univerzita obrany, Kounicova 65, 662 10 Brno, Czech Republic
Lenka Marková is an English language tester and test item developer at the Testing Department of the University of Defence Language Centre in Brno, Czech Republic.