PARTNER LANGUAGE TRAINING CENTER EUROPE
Breitenauer Strasse 16, Gebäude 253
D-82467 Garmisch-Partenkirchen

2nd Benchmark Advisory Test (BAT2)

Technical Report

March 2020

BUREAU FOR INTERNATIONAL LANGUAGE COORDINATION

PLTCE
PARTNER LANGUAGE TRAINING CENTER EUROPE

ACTFL
AMERICAN COUNCIL ON THE TEACHING OF FOREIGN LANGUAGES

Review of the 2nd Benchmark Advisory Test (BAT2)

Introduction:

Overview

The Bureau for International Language Coordination (BILC) began development of the first STANAG 6001 Benchmark Advisory Test (BAT) in 2005 as a volunteer, collaborative project: to provide an external measure against which nations could compare national STANAG 6001 test results; to promote relative parity of scale interpretation and application; and to standardize what is tested and how it is tested. In 2006, NATO's Allied Command Transformation (ACT) awarded a contract to the American Council for the Teaching of Foreign Languages (ACTFL) to operationalize and administer 4-skill BAT assessments to NATO/partner nations. Eleven nations participated in the BAT study in 2009, administering both national tests and the BAT. Overall, national test scores were higher than the BAT scores; however, the results gave national testing teams valuable information about their standardization efforts. Since the first version of the BAT, the Partner Language training Center Europe (PLTCE) and BILC have introduced the Advanced Language Testing Seminar (ALTS), the Language Standards and Assessment Seminar (LSAS), and the Faculty Development Workshop (FDW) to enhance standardization efforts even more. In 2017, PLTCE began planning for a second BAT for a more current assessment of standardization efforts.

Rationale

National STANAG 6001 tests are designed to assess an individual's unrehearsed, curriculum-independent abilities in frequently-occurring real-world communicative settings. The criticality of accurate measurement of language competence for military job requirements has imposed a strong emphasis on standardization of the STANAG 6001 testing protocols (Seinhorst). STANAG 6001 tests all follow the same basic outline and performances are judged against fixed criteria by raters who have been trained to arrive at consistent decisions. In order to obtain accurate assessment results, STANAG 6001 tests adhere to the specifications of the STANAG 6001 framework (Clifford, 2012: 54; Clifford and Cox, 2013: 51), namely that:

- each level represents a separate construct that is to be independently tested and scored
- each level is defined by a unique set of commonly occurring communicative tasks, to be accomplished in level-specific conditions, with accuracy expectations aligned with those tasks and settings.

These content, task and accuracy (CTA) expectations form the core supporting structure of both the STANAG 6001 testing system and its associated rating system.

BAT2 Purpose

The 2nd Benchmark Advisory Test (BAT2) was used by BILC- member nations in STANAG 6001-based test norming and calibration studies. Its use as a benchmark (external measure), the results of which can be compared and contrasted with the results of national tests in listening, speaking, reading, and writing, is advisory only in nature. BILC stakeholders can use data derived from comparing 21 unique national tests with the BAT2 to ***gauge the effectiveness of the community's standardization and norming efforts*** (e.g., LTS, ALTS, and various BILC-sponsored events). Likewise, individual STANAG 6001 national testing teams can use results to compare rating consistency with other national testing teams.

Contract

In December 2017, the George C. Marshall Center (GCMC) awarded a contract to ACTFL. Under the contract, in addition to administering the BAT reading and writing tests, ACTFL was to develop and administer tests in the speaking and writing skills. The contract also stipulated that a total of 210 BAT administrations were to be allocated among 21 STANAG 6001 national testing teams. Once GCMC awarded the competitive contract, GCMC became the employer, however PLTCE and BILC SMEs continued to advise on technical questions related to test item production and testing protocols. The test items developed remain under PLTCE control and cannot be used by ACTFL for other testing purposes without prior written consent by GCMC/PLTCE.

Participation

Between November 2018 and June 2019, 18 nations participated in the BAT2 project. National STANAG 6001 language testers proctored 196 BAT2 Listening, Speaking, Reading, and Writing assessments.

Literature Review

For a compilation of sources considered before, during, and after the project, see the Reading List which follows this report.

Test Development & Methodology:

Model

The BAT, as is the case with all STANAG 6001 tests, is a practical assessment in the sense that it relates to or is manifested in practice or action – it is not theoretical or ideal. It is, also, practical in that it is capable of being put to use or account. In other words, it is useful. Bachman (2007) distinguishes between language testers as researchers and as practitioners. He argues that practitioners design and develop tests that are useful for their intended purposes as compared to researchers' interest in psychological and contextual factors that

affect performance. Borsboom (2004) argues that showing test validity is about causality, as opposed to correlation. In other words, "a test is valid for measuring an attribute if variation in the attribute causes variation in the test scores." His position is that this is an ontological process in which "the attribute being measured exists and affects the outcome of the measurement procedure." He contrasts this with an epistemological process in which the researcher is concerned with investigating the reality (or existence) of the attribute. He cautions that an epistemological approach questions whether test score interpretations are consistent with a nomological network involving theoretical and observational terms. This is why Clifford (2017) calls on STANAG 6001 testers to take a scientific, not a philosophical, approach by linking an observable trait (e.g., reading ability at STANAG 6001 Level 2) to measurements of individual performance against defined Task, Condition, and Accuracy (aka Content, Task, and Accuracy) expectations. Since performance is relative to the STANAG 6001 (the criterion), not to each other, the BAT and individual national STANAG 6001 tests are Criterion-Referenced Tests. This infers that the criteria are known and real, not part of hypothesized nomological networks from which the latent trait of interest must be teased out. In his strategy to optimize a test's validity argument, Clifford specifies the alignment of three essential elements: 1) the construct to be tested; 2) the test design and its development; and 3) the scoring process. It stands to reason that all STANAG 6001 test constructs derive from the same model, in the Fulcher and Davidson (2009) sense. That is, STANAG 6001 skill level descriptors, in toto, articulate a theory of language ability circumscribing the domain of English for Interoperability Purposes. If all tests are designed to assess, not just by modality, but also by skill level, then the first two elements are aligned. If the scoring method requires testing the floor (sustained noncompensatory performance) and the ceiling (unsustained performance) within each modality's skill level, then all three elements are aligned.

STANAG 6001's overarching aim is to respond to interoperability requirements within NATO, especially as the common standard for developing language proficiency tests and recording/reporting Standardized Language Profiles (SLP). The STANAG 6001 scale was developed as a tool to describe and assess spontaneous, real-world language competence for international (military) job requirements, and is primarily intended for employers and other end-users – military commanders, personnel managers, etc. Correspondingly, the Target Language Use domain of interest articulated in STANAG 6001 could be characterized as "English for Interoperability Purposes". The standard consists of descriptions of language proficiency levels for the skills of listening, speaking, reading, and writing.

Language tests that are based on the STANAG 6001 framework are by nature proficiency tests. 'Proficiency tests' measure an individual's ability to use general, spontaneous, real-world language, regardless of the manner or the course of study in which the language was acquired. STANAG 6001 tests measure the ability to consistently complete the real-world communication tasks in the specified situations with the level of accuracy expected in those situations.

Invariably, STANAG 6001 tests are used as formal exams for various high-stakes purposes, such as employment and deployment decisions, promotions, course admission, and proficiency pay.

Design & Scoring System

During testing, each skill was assessed separately. The Reading, Listening, and Writing tests were delivered online by ACTFL affiliate Language Testing International (LTI) using tailored software. The Speaking test was administered over the telephone. Test proctors from each nation received specific information about logging onto the BAT2 website for administration. Each examinee took a unique version of the test.

Listening: The listening comprehension portion of the exam is multi-stage adaptive and covers Levels 1 through 3. The exam consists of testlets, a group of five items all at the same level, which are administered one at a time. Because the test is adaptive (relative to the ability level of the test taker), test length can range from 10 to 35 listening items. Test takers may listen to Level 1 passages twice, but only once for Levels 2 and 3. Once enough information has been gathered to assign a floor rating (the level where the examinee has demonstrated sustained performance) and a ceiling rating (the level where patterns of breakdown emerge), the test ends. To determine the final listening rating, a performance floor and ceiling must be established. The floor rating is reported on a scale of 0 to 3 which indicates the highest level of sustained ability. The ceiling is determined to be at one of three within-level rankings: Random, Emerging, or Developing and indicates the level of non-sustained performance at the next-higher level. A ceiling ranking of Random or Emerging does not result in a change to final rating; however, a ceiling ranking of Developing results in a "plus" being added to the floor rating. (See Annex A for Listening Test Specifications)

Speaking: The speaking component of the test is a multi-level adaptive test of speaking measuring proficiency from Level 1 to Level 3. The interactive speaking test consists of a variety of language tasks on a range of topics. The tester determines the test taker's highest level of sustained performance (the floor) which is demonstrated during the course of the test. Additionally, the tester establishes the level at which the test taker can no longer sustain performance (the ceiling). Test length can range from 20 to 40 minutes. One trained tester/rater tests each examinee by telephone in order to obtain evidence of the speaking proficiency of the examinee. That tester provides a preliminary rating. The digitized speaking sample of each test is saved and independently rated by an additional trained rater. The rating is based solely on the linguistic evidence demonstrated during the test. Each rater evaluates the speaking sample independently of any other raters. A third independent rating is requested in case of discrepancy between the two raters. (See Annex B for Speaking Test Specifications)

Reading: The reading comprehension portion of the exam is multi-stage adaptive and covers Levels 1 through 3. The exam consists of testlets, a group of five items all at the same level, which are administered one at a time. Because the test is adaptive (relative to the ability level

of the test taker), test length can range from 10 to 35 reading items. Once enough information has been gathered to assign a floor rating (the level where the examinee has demonstrated sustained performance) and a ceiling rating (the level where patterns of breakdown emerge) the test ends. To determine the final reading rating, a performance floor and ceiling must be established. The floor rating is reported on a scale of 0 to 3 which indicates the highest level of sustained ability. The ceiling is determined to be at one of three within-level rankings: Random, Emerging, or Developing and indicates the level of non-sustained performance at the next-higher level. A ceiling ranking of Random or Emerging does not result in a change to final rating; however, a ceiling ranking of Developing results in a "plus" being added to the floor rating. (See Annex C for Reading Test Specifications)

Writing: The writing component of the test is multi-level (1 to 3) consisting of three prompts for written responses on a variety of practical, social, and professional topics in informal and formal contexts. Prompts 1 and 2 each consist of two writing tasks: Prompt 1 has a Level 1 task followed by a Level 2 task on the same topic; and Prompt 2 has a Level 2 task followed by a Level 3 task on the same topic. Prompt 3 consists of a single Level 3 writing task. Prompts must be completed sequentially. Test takers have two hours to complete the entire test. The rating is based solely on the linguistic evidence demonstrated during the test. Examinees should provide responses to all tasks in the three prompts sequentially so that accurate ratings can be established. Each writing sample is independently evaluated by two raters. A third independent rating is requested in case of discrepancy. Level 1: Evidence demonstrated on test of sustained ability to coherently group sentences together on a simple topic using basic linking words. Level 2: Evidence demonstrated on test of sustained ability to coherently combine sentences into connected paragraphs on routine, everyday topics using appropriate vocabulary, grammar, and cohesive devices. Level 3: Evidence demonstrated on test of sustained ability to write extensive, cohesive, formal and informal texts on practical, social, and professional topics using specific vocabulary and complex grammar to convey the message accurately. The relationship and development of ideas are clear and major points are coherently organized. *Plus levels are awarded for writing ability clearly demonstrated but not fully sustained at the next-higher level.* (See Annex D for Writing Test Specifications)

Training & Norming Testers, Raters, Proctors

Norming Sessions

During the summer of 2018, PLTCE hosted two BAT2 Speaking and Writing Norming Forums. The first forum ran from 9 through 13 July; the second – from 6 through 10 August. The participants, all experienced STANAG 6001 testers, had the opportunity to become reacquainted with the BAT protocol for testing the skills of speaking and writing, participate in norming activities, conduct mock speaking tests and moderate writing prompts. Participants were encouraged to pursue certification as interlocutors or raters of BAT speaking and writing

tests. The certification process commenced after the norming forum. Participants were required to complete approximately 5 hours of web-based norming activities prior to the forum and were also expected to develop writing prompts which were, in turn, used in the BAT2 tests.

In the blended approach to the BAT2 Norming Forum sessions, instructions and specifications for creating BAT2 writing prompts were included on the ACTFL Schoology Learning Management System (LMS) as tasks for participants. Prior to the residential norming forum participants submitted writing prompts for subsequent moderation and use on the BAT2 writing test.

ACTFL facilitators collated prompts to be moderated by participants during the resident norming forum sessions. After each of the two norming forum sessions, facilitators further collated all the moderated prompts and submitted them to PLTCE for potential inclusion on the final BAT2 writing tests.

Pre-Testing Writing Prompts

PLTCE teamed with US Defense Language Institute English Language Center (DLIELC) to pre-test the BAT2 writing test prompts due, primarily, to the availability of relatively large numbers of international test takers, many from NATO partner nations. PLTCE prepared complete, fully formatted copies of three BAT2 writing tests, each with an accompanying questionnaire, to send to DLIELC for pre-testing. Using the Question Mark language-testing platform, DLIELC tested 74 international students with three trial versions of the BAT2 writing test. Following the test, each test taker had the opportunity to provide qualitative feedback on the writing prompts, test format, and testing experience.

DLIELC transferred important metadata about the test-taking population, the writing test samples, and qualitative feedback to PLTCE. In return, PLTCE provided individual diagnostic feedback on the writing samples and transferred to DLIELC certificates of appreciation to individual test takers. Subsequent analysis of pre-testing quantitative and qualitative data indicated that the three sets of writing test prompts could be used as part of the BAT2 battery of tests.

PLTCE, then, sent the writing prompts to ACTFL so LTI personnel could add them to the BAT2 delivery system. By November 2018, LTI had the BAT2 writing test prepared for live delivery.

Rater & Tester Certification

All participants in the Rater Norming Forums had the opportunity to apply for certification as an official BAT2 tester and/or rater. A total of six ACTFL-certified BAT2 speaking testers and seven speaking and/or writing raters conducted and rated the BAT2 productive skills assessments. Testers came from Croatia, Denmark, Georgia, Latvia, Norway, and Romania – raters from Bosnia & Herzegovina, Bulgaria, Czech Republic, Lithuania, North Macedonia, Slovakia, and Slovenia. It is significant that all testers and raters were professional STANAG 6001 language testers and graduates of the BILC-sponsored Advanced Language Testing Seminar (ALTS).

Tester & Rater Liaison with ACTFL

Upon certification, testers and raters coordinated with Language Testing International (LTI) sourcing specialists for onboarding and scheduling details. Throughout the period November 2018 – June 2019, BAT2 testers and raters informed LTI of their availability for testing and/or rating.

Proctor Selection / Training

The primary roles of the BAT2 proctors were to safeguard the integrity of the test and ensure that the test was administered fairly and consistently. Proctors checked to make sure that computer hardware and software were set up correctly and functioning properly. Before testing, they verified the identity of each BAT2 examinee, issued passkeys, and assisted with logging on to the testing system. During the test, proctors monitored examinees to maintain a quiet setting and to safeguard against test item compromise. All BAT2 proctors received Proctor Instructions and signed/returned to PLTCE a BAT2 Proctor Agreement.

Pre-Viewing the BAT2 Reading and Listening Tests

Before the official launch of the BAT2 on 19 November 2018, all Norming Forum participants had the chance to take the BAT2 Reading and Listening tests. Most of the Norming Forum participants later acted as national BAT2 POCs or test proctors, so taking the tests gave them a good idea of what their test takers could expect and prepared them for fielding test taker questions.

Pre – launch checks

Scheduling and Coordinating BAT2 Tests

Scheduling and coordinating BAT2 tests occurred in two phases. Phase 1 involved, mostly, information sharing with BAT2 national POCs and proctors. PLTCE requested that each participating nation complete a Questionnaire for BAT2 Participating Nations via a Google form link. Proctors received a BAT2 Proctor Agreement, which they signed and returned to PLTCE, and instructions for facilitating BAT2 administration. Additionally, each nation got a copy of the BAT2 Examinee Guide and a link to online BAT2 demo tests. Phase 2 required that participating nations identify examinees (normally, a total of ten) and submit national STANAG 6001 SLPs to PLTCE (if tested in advance).

Testing Platforms

LTI delivered via internet all BAT2 Reading, Listening, and Writing tests using their own language testing website, plus BYU's proprietary algorithm for Reading and Listening. In order to facilitate communications between BAT2 Speaking testers and test takers, who were geographically separated, LTI scheduled all Speaking tests as telephonic OPIs. LTI schedulers set

up a direct-dial system by issuing international calling cards to participating nations to defray the costs of the extended phone calls.

Meeting Technical Requirements in Country

Due to the online delivery of the BAT2 Reading, Listening, and Writing assessments, national POCs/proctors had to ensure they could meet LTI's technical requirements for computer-delivered testing. One of the initial hurdles was the inability to use the Internet Explorer browser to load the tests. Generally, Google Chrome and Mozilla Firefox were good alternatives.

Choosing Test Takers

PLTCE advised national POCs to select ten test-takers with valid STANAG 6001 SLPs no older than six months in order to draw a fair comparison between national tests and BAT2. Because most NATO assignments require more than Level 1 proficiency, checking for alignment at Levels 2, 2+, and 3 was the focus of this round of testing. Based on the recommendation of the BILC senior advisor, most national BAT2 POCs selected participants whose speaking skills were at one of these proficiency levels. Proctors made available a BAT2 Examinee Guide and a link to online demo tests to each of the examinees well before their scheduled BAT2 tests.

Preparing Test Takers

The BAT2 Examinee Guide familiarized test takers with the skill components of the BAT2, including test format, topical content, text types, task and accuracy requirements, administration procedures, scoring procedures, and, importantly, sample test items.

BAT2 Launch

Liaison with ACTFL

PLTCE, BAT2 POCs/proctors, and testers/raters all worked closely with ACTFL in the closing days (and hours) before launching BAT2 in order to settle remaining technical and scheduling issues.

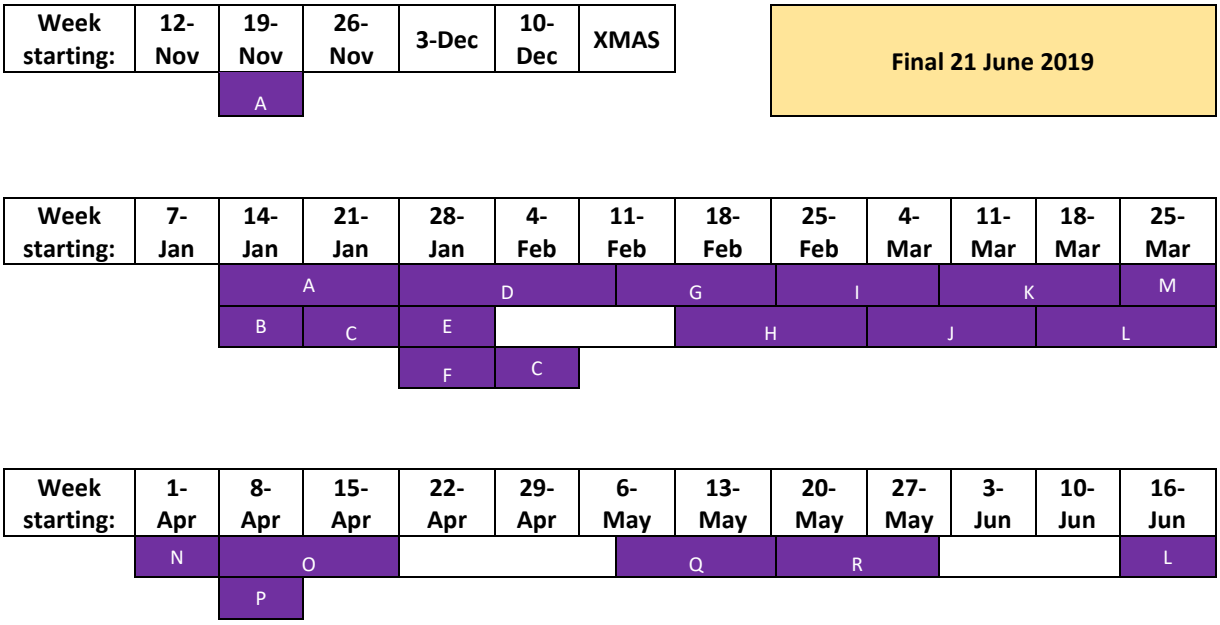
Registration of Test Takers

By November 2018, PLTCE had produced a tentative testing schedule for the 18 participating nations. One nation tested in mid-November 2018, while the remaining 17 scheduled their test for January through June 2019. Generally, about two weeks before scheduled testing for each nation, PLTCE sent individualized access codes to proctors so test takers could take the BAT2 Reading, Listening, and Writing assessments during the nation's scheduled testing period.

Communications

Since the BAT2 testers conducted Speaking tests telephonically, LTI issued telephone calling card numbers to each participating nation to cover the costs of the international calls, all of which were routed through LTI's New York office. LTI performed communication checks with proctors before the start of Speaking testing in each nation.

Test Administration



For nations testing in one week, proctors, generally, administered Reading, Listening, and Writing tests during the first two days; then, LTI scheduled Speaking tests during the last three days. For those testing in two weeks, proctors held Reading, Listening, and Writing tests during Week 1, while LTI scheduled Speaking tests during Week 2.

Collecting Qualitative Data

As nations finalized BAT2 assessments, participants completed questionnaires about their test taking experiences and perceptions (see Annex F). Individuals submitted all survey data directly to PLTCE for qualitative analysis. Additionally, proctors completed questionnaires related to their own experiences and perceptions.

Reporting Results

PLTCE acted as a clearinghouse for all BAT2 scores. Upon completion of all BAT2 assessments for each nation (i.e., 10 sets of scores), PLTCE released score reports to national BAT2 representatives. PLTCE, also, reported qualitative data back to individual nations. To recognize

BAT2 project participation, PLTCE produced individual certificates of appreciation for each individual test taker. Digital copies were forwarded to the national BAT2 action officers.

Values & Consequences

Among the issues related to BAT2 score use considered during the project:

- ✓ The score on the test is an adequate reflection of the observed test behavior.
- ✓ The assessment yields results that are consistent across assessment contexts.
- ✓ The assessment provides information on test takers' English-language proficiency that is consistent with content, task, and accuracy statements in STANAG 6001 skill level descriptions. The test tasks are adequate proxies for those performed in the multinational military interoperability domain.
- ✓ Score-based decisions are appropriate, well communicated, and function as benchmark scores for national testing teams and other BILC community stakeholders to use for calibration and norming studies.
- ✓ The consequences of using the BAT and the decisions informed by the BAT are beneficial to all stakeholders.

BAT2 Analysis & Lessons Learned:

From the beginning of the project, PLTCE intended to bring together a group of language testing SMEs to analyze BAT2 results (relative to the national STANAG 6001 SLPs of project participants), review lessons learned, and assess the validity, reliability, usefulness, and fairness of any findings. PLTCE invited four language testers to meet in Garmisch from 24 through 28 June 2019. Organizers and participants agreed on pre-meeting coordination and the basic meeting agenda.

Pre-Meeting Coordination

Before actually meeting, the SMEs shared all available data such as Speaking and Writing tester/rater data and Listening and Reading item (bank) data. Group members updated test specs to reflect the actual BAT2 assessments administered. Specifically, SMEs looked at how to analyze Listening and Reading data for multistage adaptive testing and how to calculate inter-rater reliability (IRR) and/or rater agreement for Speaking and Writing tests.

Topics, Analyses, and Discussions

After a briefing on the project and meeting goals, SMEs discussed approaches to analyzing BAT2 data. Examination of the quantitative data included a comparison of BAT1 (2009) and BAT2 (2019), as well as BAT2 vs. (then) current national STANAG 6001 SLPs. Qualitative data included test-taker, proctor, and tester/rater feedback. Looking specifically at Speaking and Writing, members reviewed IRR and rater agreement stats. All agreed that the 2018 norming forums had a very positive effect on the BAT2 cooperative effort and should be repeated in the future,

whenever possible. Analysis of the Reading and Listening test data suggested that the novel MST format, individual item-response timing, and questionable text/passage quality could have had a negative impact on overall BAT2 receptive skill testing.

Conclusions and Recommendations

All participants agreed that the BAT2 project demonstrated that the BILC language testing community's standardization and norming efforts, especially over the past ten years, have had a positive effect on STANAG 6001 testing. These efforts include:

- Development of an Advanced Language Testing Seminar (ALTS)
- Sponsoring of an annual STANAG 6001 testing workshop
- Enhancement of the BILC website's testing resources page
- Initiation/expansion of bilateral/regional cooperation
- Improved networking with colleagues and international testing experts
- Sharing of STANAG 6001 testing best practices
- Drafting of a STANAG 6001 Roadmap to Validity
- Conducting various BILC assistance visits
- Research and academic achievements of STANAG 6001 testers

The group also recognized the importance of involving national STANAG 6001 language testers in the development, trialling, and administration of community-wide benchmark assessments. For example, PLTCE-hosted tester/rater norming forums resulted in significant gains in BAT2-National STANAG 6001 testing correlations. Pre-testing of BAT2 Writing prompts at DLIELC contributed to the strength of the Writing test, which went from the assessment with the least amount of correlation with national assessment to the one with the most (ahead of Listening, Speaking, and Reading). By contrast, the group agreed that not involving the BILC community in the development of BAT2 Reading and Listening assessments resulted in decreased correlations with national tests.

Additionally, a tremendous response to post-BAT2 questionnaires revealed some important lessons learned among test takers, proctors, and testers/raters. The testers and raters remarked on the many benefits derived from the rater norming forums, not just for individuals but, also, for national testing organizations. Despite overall results to the contrary, most test takers considered the BAT2 assessments easier than their national tests. Many respondents expressed a preference for computer-based testing (as opposed to paper-based national tests). Nevertheless, there were numerous concerns raised about technical difficulties experienced during the online BAT2 Listening, Reading, and Writing assessments.

Participants, also, began work on identifying Speaking and Writing samples to use at the next STANAG 6001 Testing Workshop and to replace outdated materials for the LTS and ALTS.

Finally, the group looked ahead to the 2019 STANAG 6001 Testing Workshop on how to present results, highlight lessons learned, and deliver meaningful workshops reinforcing lessons learned and best practices from the BAT2 project.

Follow-through and Look Ahead

STANAG 6001 Testing Workshop

During the 2019 Testing Workshop in Tours, France, BILC language testers delivered presentations, hosted panel/hot topic discussions, and conducted workshops, predominantly related to BAT2 lessons learned. In particular, BILC language testers agreed that it is important to run Speaking & Writing rater norming sessions before large-scale testing events, to use blended/hybrid learning to increase exposure to norming materials and samples, and to collect test taker/proctor/rater feedback to conduct qualitative analysis.

Reaping the Rewards of the BAT2 Project

The BAT2 project yielded a wealth of new resources to be added to the BILC website and to be used at future Testing Workshops. A special working group (BAT2 testers/raters) convened at PLTCE to select new Speaking and Writing samples to use during the ALTS and for other BILC language testing purposes. Recommendations from national STANAG 6001 language testing teams included: conducting norming sessions in conjunction with the annual Testing Workshop; collaboration on a shared reading item bank; and training on converting from paper-based to computer-based testing.

BAT2 Timeline

<table>
<tr><th></th><th>Jan/Feb</th><th>March</th><th>April</th><th>May</th><th>June</th><th>July</th><th>Aug</th><th>Sep</th><th>Oct</th><th>Nov/Dec</th></tr>
<tr><td>2017</td><td colspan="3">Establishment of participant nations, goals and project objectives</td><td></td><td></td><td colspan="5">PLTCE competes and awards contract for BAT2 administration & norming/certification of BAT2 testers & raters</td></tr>
<tr><td>2018</td><td colspan="5">Planning and coordination with nations & ACTFL contractor</td><td colspan="2">Norming Forums x2 (July & August)</td><td>Trial Writing Tests
Certify Testers/Raters</td><td colspan="2">Launch Live BAT2</td></tr>
<tr><td rowspan="2">2019</td><td colspan="5">BAT2 registration, administration & score reporting</td><td colspan="5">After action</td></tr>
<tr><td colspan="4">Continue Live BAT2</td><td>End of Testing</td><td colspan="2">Consolidation & preparation of reports / lessons learned</td><td colspan="3">Report BAT2 lessons learned at BILC STANAG 6001 Testing Workshop in Tours (September).</td></tr>
</table>

		BAT2 Analysis Meeting		Download/file BAT2 audio files and test scripts from ACTFL to GCMC servers Modifications & updates to LTS/ALTS curricula (Jan-Mar 2020)

Acknowledgements

PLTCE is extremely grateful for the enthusiastic participation and enormous contributions made by ACTFL/LTI staff, participating nations (specifically, test takers, proctors, and POCs), Speaking and Writing tester/raters, subject matter experts, and significant contributors to the BAT2 project.

Participating Nations

Austria, Belgium, Bosnia & Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Georgia, Hungary, Italy, Latvia, Lithuania, North Macedonia, Norway, Romania, Slovakia, Slovenia, and Spain.

Testers and Raters

Testers: Martina Aleric, Birgitte Grande, COL Corina Ispas, Julija Kolosovska, Allan Kristiensen, and Tamar Shavlakadze

Raters: Tadeja Hafner, Irena Katauskiene, Krassimira Koleva, Jan Krivka, Major Nermin Nuhic, Gabriela Repova, and Vladimir Trpkoski

Subject Matter Experts

- Dr. Ray Clifford, Brigham Young University
- Dr. Troy Cox, Brigham Young University
- Mary Jo Dibiase Lubrano, Yale University
- Peggy Garza, PLTCE
- Roxane Harrison, PLTCE
- David Oglesby, PLTCE
- Dr. Elvira Swender, past Director ACTFL Professional Programs

Significant Contributors

- COL Corina Ispas – the only BAT1 and BAT2 tester; facilitator during BAT2 review in Tours, leader of the new Speaking and Writing sample selection process.
- Dr. Edelmira Nickels – coordinated and oversaw administration of BAT2 Writing pre-testing at the Defense Language Institute English Language Center.

Reading List

ALTE (Association of Language Testers in Europe). (2011). Manual for Language Test Development and Examining. Council of Europe: Language Policy Division. Retrieved 9 July 2017, from www.coe.int.

Anastasi A. (1950). The concept of validity in the interpretation of test scores. Educational Psychology Measurement, 10, 67–78.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association

Bachman, L. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment, in: Fox, J. et al (ed.), Language Testing Reconsidered. Ottawa, University of Ottawa, 41–71.

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

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

Chapelle, C. A. (1999). Validation in language assessment. Annual Review of Applied Linguistics, 19, 254–272.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). Building a validity argument for the test of English as a foreign language. New York, NY: Routledge.

Chapelle, C. A., Enright, M. K., &. Jamieson, J. (2010). Does an argument-based approach to validity make a difference? Educational Measurement: Issues and Practice, 29(1), 3–13.

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

Clifford, R.T. (2001). Opening Remarks. BILC 2001 Conference Report (Segovia, Spain), 17–39.

Clifford, R.T. (2012). It is Easier to Malign Tests than it is to Align Tests, in: Tschirner, E. (ed.), Aligning Frameworks of Reference in Language Testing. Tübingen, Stauffenburg Verlag, 49–56.

Clifford, R.T. & Cox, T.L. (2013). Empirical Validation of Reading Proficiency Guidelines. Foreign Language Annals, 46, 1: 45–61.

Clifford, R.T. (2017). Developing language proficiency tests with more evidence and less inference. Presentation at Defense Language Institute Foreign Language Center (DLIFLC), Monterey, CA.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281–302.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179–197.

Embretson, S. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. Psychological Methods, 3, 380–396.

Fulcher, G. & Davidson, F. (2007). Language testing and assessment: An advanced resource book. New York, NY: Routledge.

Fulcher, G. & Davidson, F. (2009). Test architecture, test retrofit. Language Testing, 26 (1): 123–144.

Kane, M. (1992). An argument-based approach to validation. Psychological Bulletin, 112, 527–535.

Kane, M. (2006). Validation. In R. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013), Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50: 1–73.

Kenyon, D. M. & MacGregor, D. (2016). Standard Setting Workshop at the 38th annual meeting of the Language Testing Research Colloquium (LTRC), University of Palermo, Palermo, Italy.

Lissitz, R., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. Educational Researcher, 36, 437–448.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, Monograph Supplement, 3, 635–694.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed. pp. 13–103). New York, NY: American Council on Education and Macmillan.

Mislevy, R., Steinberg, L., &. Almond, R. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1, 3–62.

Seinhorst, G. (2017). BILC policy recommendations on the portability of STANAG 6001 Language Certifications. Retrieved 2 February 2018, from www.natobilc.org.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), Review of research in education (pp. 405–450). Washington, DC: American Educational Research Association.

Sireci, S. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), The concept of validity (pp. 19–38). Charlotte, NC: Information Age Publishers.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), Review of research in education (pp. 405–450). Washington, DC: American Educational Research Association.

Weideman, A. (2012). Validation and validity beyond Messick. Per Linguam, 28(2): 1-14.

Weideman, A. (2013). Applied linguistics beyond postmodernism. Acta Academica, 45(4): 236-255.

Zumbo, B. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. Lissitz (Ed.), The concept of validity (pp.65–82). Charlotte, NC: Information Age Publishers.

Annex A

NATO BILC
Benchmark Advisory Test 2 (BAT2)
Listening Test Specifications

Purpose of the BAT2 Tests	The 2nd version of the Benchmark Advisory Test (BAT2) is used by Bureau for International Language Coordination (BILC) member nations in STANAG 6001-based test norming and calibration studies. Its use as a benchmark (external measure), the results of which can be compared and contrasted with the results of national tests in listening, speaking, reading, and writing, is advisory only in nature. BILC stakeholders can use data derived from comparing 21 unique national tests with the BAT2 to gauge the effectiveness of the community's standardization and norming efforts (e.g., LTS, ALTS, and various BILC-sponsored events). Likewise, individual STANAG 6001 national testing teams can use results to compare rating consistency with other national testing teams.
Construct Definition	One of the definitions of proficiency relates to a general type of knowledge or competence in the use of a language, regardless of how, where or under what circumstances it has been acquired. Proficiency is conceptualized here as a global construct that transfers across contexts, tasks, and events, while proficiency tests attempt to sample the underlying competence by eliciting behaviors on tests that generalize to domains of interest. Proficiency can also be defined as the language knowledge that is needed to function in a future situation. The performance elicited in a proficiency test is usually measured or judged against a set of criteria, represented in a rating scale. The BAT2 will measure proficiency in listening comprehension in accordance with STANAG 6001. Listening comprehension will be measured in accordance with STANAG 6001 as the ability to: • process extended samples of realistic spoken language, automatically and in real time • understand the linguistic and cultural information that is unequivocally included in a spoken passage • make any speaker-intended inferences that are unambiguously implied by the context/content of the spoken passage.
Delivery of BAT2 Listening Test	Computer-based (online) delivery. This test will be administered and scored online using a tailored software product.

Test Format	The listening comprehension portion of the exam is adaptive and covers Levels 1 through 3. The exam consists of testlets, a group of five items from the same level, which are administered one at a time. Each testlet is constructed so that it has the same difficulty as the other testlets at its level. The testlets function as mini-tests, and examinee performance on each testlet results in one of three outcomes: 1) the examinee was able to sustain listening performance at the level; 2) the examinee was unable to sustain performance at the level; or 3) more information is needed. Depending on examinee performance, the examinee receives another testlet from the same level, a testlet from the higher level, or a testlet from the lower level. Once enough information has been gathered to assign a floor rating (the level where the examinee has demonstrated sustained performance) and a ceiling rating (the level where patterns of breakdown emerge), the test will end. The minimum number of testlets in a given exam is two (or 10 items) and the maximum is seven (or 35 items) with the length of the test ranging from six to 60 minutes. The exam items are multiple-choice that are timed by level and each consists of a brief orientation, a stem, and five options (or choices). The fifth option is always "I don't know," and is included so examinees are not forced to guess when the item is above their ability level. All test materials, examples, and instructions are in English. *An example multiple-choice exam item:* A voicemail message – (representing the orientation) Transcript You have reached the State College switchboard. The office hours are from 8:00 a.m. to 5:00 p.m. Monday through Friday. If you wish to leave a message, please wait for the tone, and then give your name and telephone number. Thank you. What does this voice mail greeting tell us? **(A) How to leave a message. (key)** (B) To wait for the operator. (distractor) (C) How to make an appointment. (distractor) (D) The college's telephone number. (distractor) (E) I don't know. (option) After submitting each item for scoring, the examinee cannot return and change the item answer. If the examinee takes longer than the time allotted for a specific item, the option "I don't know" will be automatically selected and the test will proceed to the next item.

Topical Content	Content areas for all listening items are targeted to the general listener. Topics at all levels may include texts that are relevant to NATO operations. Lower level content areas will include everyday survival and work-related topics. At higher levels, content areas may include: • military and security issues • economic and political matters • scientific and technical issues • cultural and social issues.
Spoken Texts	The BAT will focus on authentic texts (spoken by native speakers for communicating with native listeners), taken directly from American, British, Canadian, and other authentic sources, with good acoustical quality. The pronunciation of the native speakers needs to be representative of a mainstream variety of the aforementioned "Englishes." Texts will be selected to include a variety of speakers, talking at a rate that is normal and realistic for the text type. Monologues, dialogues, debates and interviews are some examples of text types. Texts may include redundancies, false starts, fillers and other features particular to spoken language. Overlapping speech should be avoided at all levels. A small number of semi-authentic texts may be used as needed. More specifically, level one texts may represent a combination of texts derived from authentic sources, as well as a number of dialogues based on semi-scripted topics. All texts should be self-standing and representative of the target level

Types of Spoken Passages	***Level 1*** Short, discrete, simple announcements/exchanges: weather reports; broadcasts of sport scores; announcements at public events; emergency announcements; introductions to TV and radio programs. These passages are not linguistically complex and deal with common, everyday situations requiring listening skill. The spoken texts generally contain very basic vocabulary. ***Level 2*** Straightforward, detailed information of events occurring in multiple time frames: instructions or orders; short factual news broadcasts; factual narrations/descriptions in a news broadcast; short concrete conversations; longer telephone messages. These passages deal with factual occurrences in the everyday world. They contain concrete factual vocabulary and may include some linguistically complex structures. These texts have an organization that is predictable for the target language. ***Level 3*** Extended, abstract discourse with complex syntactic structures: interviews on social, scientific, or political issues; broadcast editorials; speeches or lectures; debates and discussions; recorded meetings, conferences, or briefings; more complex conversations on TV and radio programs. These spoken passages demonstrate a wide variety of discourse structures and a wide range of vocabulary. They contain complex argumentation, including hypothesis, supported opinion, analysis, implications and some nuances. Main ideas are often not stated explicitly and require "listening between the lines."
Tasks and Accuracy	Listening tasks and accuracy requirements are in accordance with NATO STANAG 6001.

Test Administration	Before examination, the identity of each examinee is checked and verified. On-site proctors assist each examinee in logging onto the examination website and supervise the examination environment. All test instructions appear in English. Examinees are instructed to: • give electronic devices such as mobile phones, cameras, smartwatches, and other items to the proctor for the duration of the test; • look only at their computer screen and not allow others to see their screen; • report any suspicious activities; • avoid talking with others; • and refrain from consulting outside resources, such as dictionaries or web pages. In addition, examinees are instructed not to disclose the contents of the test to anyone, including, but not limited to, teachers, employers, or friends. While each item is timed, the examinee may move to the next item before the time has expired. Bi-level scores are reported to the POC for the country in the floor-ceiling format. The floor rating is reported on a scale of 0 to 3 which indicates the sustained ability level. The ceiling is reported as one of three levels: Random, Emerging, or Developing and indicates the level of performance at the next level.
Validation	The initial item validation work and setting of cut scores was accomplished using a modified Angoff rating procedure and Rasch analysis.
END	

Annex B

NATO BILC
Benchmark Advisory Test 2 (BAT2)
Speaking Test Specifications

Purpose of the BAT2 Speaking Tests	The purpose of the BAT2 is to provide nations with an external criterion for validation of their general proficiency speaking tests based on NATO STANAG 6001. This test is designed for NATO and partner nation military and civilian personnel who are non-native speakers of English. The BAT2 speaking test assesses general English language speaking proficiency up to STANAG 6001 Level 3, regardless of how it was acquired.
Construct Definition	One useful definition of proficiency focuses on general competence in the use of a language, regardless of the specific circumstances in which that language was acquired. That is, proficiency is not related to a particular curriculum, training course, set of materials, or institution. Based on this definition, a proficiency test attempts to provide a sufficiently large and varied sample of language tasks to demonstrate what examinees are able to do in that language. The aim is to measure an underlying competence which can then be generalized to similar domains of interest. The BAT2 is a criterion-referenced test that will measure speaking proficiency in accordance with the STANAG 6001 speaking scale. That scale, with its descriptors of the tasks, content, text type, and accuracy required for each speaking level, will provide the criteria for rating examinees.
Definition of speaking	Speaking proficiency is defined as active, automatic, use of one's internalized language and culture expectancy system to efficiently and purposefully communicate spoken language in a variety of interactional and transactional unrehearsed tasks according to STANAG 6001 speaking level descriptors for levels 1, 2 and 3.
Administration and rating of the BAT2 Speaking Test	The BAT2 speaking test will be scheduled online and administered telephonically under a contract with the American Council on the Teaching of Foreign Languages (ACTFL). A trained and certified tester will test each examinee by telephone in order to obtain a ratable sample of speaking proficiency. The sample will be recorded using a digital voice recording system that is stored on a secure Internet data-base. The tester who conducts the interview will provide an initial rating. The digitized speech sample of each test will be independently rated by an additional trained and certified tester/rater. In cases of a discrepancy between the first two ratings, a third trained rater will rate the speech sample independently. A final rating will be assigned when two ratings agree. The BAT2 is based on the principles of BILC approved Best Practices for STANAG 6001 language testing. This test is a multi-level, adaptive test of speaking proficiency up to and including Level 3 of STANAG 6001. The BAT2 is based on the principles of BILC approved Best Practices for STANAG 6001 language testing.

Elicitation technique overview	The tester will elicit a speaking sample that demonstrates the highest level at which the speaker can sustain all of the criteria for the level (floor) and the level at which the speaker can no longer sustain the criteria (ceiling) during the test. The tester will require that the examinee perform a variety of language tasks on a variety of topics appropriate for the working level of the test. Probes will be used to determine an examinee's ability to perform at the next higher level. If the next level can be sustained, then all the requirements for that level will be tested. Examinees who fully meet the Level 3 proficiency requirements will not be further probed for higher level performance. Level 3 is the highest level that can be assessed by the BAT2.
General prompt information	The BAT2 for speaking is an interactive and adaptive assessment between a certified tester and the individual whose proficiency is being evaluated. Prompts will reflect the tasks outlined in the STANAG 6001 speaking descriptors for levels 1, 1+, 2, 2+ and 3.
General task and topic information	The speaking test will include a minimum number of tasks on at least five topic areas at the estimated level assigned by the rater (working level). The speaking test will include at least one role-playing situation that establishes the ability to successfully handle the role play for the level. See individual levels for lists of appropriate topics.
Test Length	The length of the test will range from 20 to 40 minutes depending on the level and the complexity of obtaining a ratable sample.
Test Phases ***Illustrated by table in annex 1***	The test will consist of three phases. Each phase of the exam is designed to check the tasks, content areas, text type, and expectations of accuracy of the level(s) that seem to apply to the examinee and to elicit spoken language that can be rated against the STANAG 6001 descriptors. A brief description of each of the parts and their purpose(s) appears below: **Initial Phase** ▪ The first part of the test is designed to help raters establish the working level of the examinee. Strategies for conducting this phase are presented in the tester training. (Testers will be trained to be as efficient as possible; however, examinees will not be penalized if the tester's initial assessment of the working level is inaccurate.) ▪ Sustained or unsustained performance of any required task during the initial phase will not be considered as a part of the core test. For example, if a candidate proves a sustained paragraph length past narration during the Initial Phase, that performance will only help to determine the opening working level of the Core Test. During the core test, the tester will need to elicit that function again to fulfill the requirements of L2. **Core Test** ▪ The core test presents the examinee with the opportunity to produce spoken language that best represents the examinee's level (level checks) as well as the opportunity to show the limit of their ability (probes). Within the core test these level checks and probes will be interwoven as appropriate. This includes

having two probes (different tasks and different topics) in order to show the limit of ability at the next higher level.

- The core test will include a role play situation that checks the examinee's ability to handle a situation. There must be a role-play at the working level of the candidate. For example, if the final rating is a 2+, then the interview must contain a L2 role-play.

Final Phase

The final phase of the test serves as a transition phase to end the test. It is rarely used as an opportunity to elicit level checks and never used to probe.

	Initial Phase	**Core Test**		**Final Phase**
Phases / Perspectives	**Warm-up**	**Level Check**	**Probes**	**Wind Down**
		Iterative Process		
Psychological	Relaxes examinee	Proves to examinee what he or she can do	Proves to examinee what he or she cannot do	Returns examinee to level at which he or she functions most accurately
Linguistic	Reacquaints examinee with language if necessary	Checks for tasks and content which examinee performs with greatest accuracy	Checks for tasks and content which examinee performs with least accuracy	Chance to check that the iterative process is complete
Evaluative	Provides testers with preliminary indication of level of speech skills	Finds the examinee's speaking level	Finds level at which examinee can no longer speak accurately	Confirms global rating; no new information

Ratable Sample

A speech sample will be considered "ratable" when the above-mentioned conditions and requirements have been met according to the STANAG 6001 descriptors. In the BAT2 speaking test the tester will elicit a speaking sample that demonstrates the floor (the level of sustained performance) and ceiling (the level at which performance is no longer sustained) of the examinee's proficiency at the time of the test. A ratable sample is a sample of speech that contains:

- All three test phases.
- Evidence of a speaker's ability to perform the tasks of a level across the content areas for the level and meeting the expectations for accuracy and text type of that level
- Evidence that the speaker cannot consistently perform at the next higher level
- All required tasks at level.

	▪ A minimum of five topics within the core test.
Rating Protocol	The examinee's performance will be rated globally by using the STANAG 6001 speaking descriptors and the rating factor grid based on the full descriptors. The rating will be based solely on the linguistic evidence demonstrated during the test. Scores will be reported as: 0+, 1, 1+, 2, 2+ and 3.
Test Level 1 Content, tasks and accuracy	▪ The examinee needs to perform required tasks for the level. There are 3 required tasks for Level 1. These tasks are: ask and answer simple questions related to daily life, engage in short conversations, and handle basic survival situations in a role-play. ▪ Discourse is sentence-level.
Test Level 2 Content, tasks and accuracy	▪ The examinee needs to perform required tasks for the level. There are 5 required tasks for Level 2 on a minimum of 3 topics. These tasks are: narrate in the past, report on a current event, provide a physical description, give instructions and/or directions, and handle a familiar situation with complication in a role-play. ▪ There must be at least two probes for L3 speech. ▪ Discourse is paragraph-length when the task and topic require it. ▪ Good control of appropriate verb forms for past, present, and future time frames is expected, although there may be some errors.
Test Level 3 Content, tasks and accuracy	▪ The examinee needs to perform required tasks for the level. There are 4 required tasks for Level 3 on a minimum of 3 topic areas. These tasks are: support an opinion, convey an abstract concept, hypothesize, and resolve a problem in an unfamiliar situation in a role-play. ▪ An L3 role-play is required. ▪ Discourse is extended cohesive discourse when the task and topic require it. ▪ Grammatical control is sufficient to discuss topics appropriately in both formal and informal speech, even though there may be occasional errors that do not distort the meaning.
Base Levels and Plus Levels	Plus ratings will be the result of the elicitation process consisting of interwoven level checks and probes, i.e. they will be derived from a clearly established floor of performance, as well as partial success at the probed level (performance that is more than half way to the next level). Testers should avoid inferring the plus ratings based solely on candidates' inability to sustain performance at the next higher level.

Level	Tasks / Functions ***What a person can do with the language*** *(tasks accomplished, attitudes expressed, tones conveyed)*	Content / Topics ***What a person can talk about*** *(subject areas, activities and jobs addressed; settings)*	Accuracy ***How well a person can use the language*** *(acceptability, quality and correctness of message conveyed)*	Text produced *(length and organization of utterance; kinds of discourse)*
0	No functional ability.	None or isolated words.	Unintelligible.	Random words and phrases.
0+	Can make short utterances and ask very simple questions using memorized material and set expressions.	Immediate survival needs such as greetings, brief personal data, numbers, time expressions, common objects.	Understandable with difficulty even to a native speaker used to dealing with foreigners.	Memorized words and short phrases.
1	Can create sentences; begin, maintain, and close short conversations by asking and answering simple questions; satisfy simple daily needs; resolve basic situations.	Everyday survival topics and courtesy requirements.	Intelligible with some effort to a native speaker used to dealing with foreigners; often miscommunicates.	Discrete sentences.
1+	Able to participate in predictable conversations about all survival needs and limited social demands; shows limited/inconsistent ability to describe, narrate, give instructions.	Basic needs, own back-ground, family, interests, travel, and simple work-related matters.	Faulty but comprehensible to a native speaker used to dealing with foreigners.	Strings of related sentences.
2	Can describe people, places, and things; narrate current, past and future activities in full paragraphs; state facts; give instructions or directions; ask and answer questions in the workplace; deal with non-routine daily situations.	Concrete topics such as own background, family, interests, work, travel, and current events.	Understandable to a native speaker *not* used to dealing with foreigners; sometimes miscommunicates.	Full paragraphs, minimally cohesive.
2+	Able to communicate in many informal and formal conversations; uses language effectively to describe, narrate, report facts, give detailed instructions and directions, handle unfamiliar situations; uses it less effectively to support opinions, clarify points, answer objections.	Practical, social, everyday professional topics, particular interests, special fields of competence, and to some extent abstract topics.	Communicates relatively well with native speakers not used to dealing with foreigners. Speech is usually appropriate to the situation, with occasional errors in vocabulary, more complex structures, or pronunciation.	Discourse beyond the paragraph level.
3	Can converse in most formal and informal situations; discuss abstract topics; support opinions; hypothesize; deal with unfamiliar topics and situations; describe in detail; clarify points.	Practical, social, professional and abstract topics, particular interests, and special fields of competence.	Speaks readily, with only sporadic non-patterned errors in basic structures. Errors almost never interfere with understanding and rarely disturb the native speaker.	Extended discourse.

STANAG 6001 - Base and Plus Level Tasks, Content, Accuracy and Text Produced Table

Rating Protocol	

	Raters will submit their ratings to ACTFL for determination of a final score and for formal notification through approved channels. Neither the score nor any other information about the test will be discussed directly with the examinee. The following procedures will be followed for rating the speech sample: 1) All samples will be rated by 2 independent certified testers/raters. 2) Tester elicits a ratable sample over the telephone. 3) Tester reviews the sample, and assigns a rating. 4) Tester submits the rating to ACTFL through approved channels. 5) ACTFL assigns sample to a different rater. 6) Rater listens to the sample and submits an independent rating to ACTFL through approved channels. 7) If ratings are identical (base and plus level), ACTFL assigns an official score, and reports it to appropriate authorities (including PLTCE) through approved channels. 8) If there is a rating discrepancy, it will be arbitrated by further independent ratings.
Quality Control of Rating Reliability	ACTFL will manage all quality control procedures. On an on-going basis, twenty percent of all speaking tests will be randomly selected for quality control of structure, elicitation techniques and rating. All reports on results and any other information of interest will be provided to PLTCE. If PLTCE determines it is necessary to change the percentage of tests reviewed or the frequency of reporting, this change will be made.
Testers and Tester Training	The training will combine face-to-face instruction and online practice. Testers/raters will be required to demonstrate the ability to consistently elicit ratable samples and rate them with a high degree of accuracy and inter-rater reliability. An important aspect of the tester training will be the interpretation and understanding of the NATO STANAG 6001 scale and the application of this standard to operational testing. Testers will be trained to control the test process and content while allowing the examinee to show his or her language ability at its best.
Tester/Rater Criteria	Tester refers to the person who conducts and rates the test. Rater refers to the person who rates the test. The tester is always the first rater. The testers and raters may be native speakers or non-natives whose own speaking level meets or exceeds Level 3. All testers and raters must be certified by ACTFL or other organizations designated by PLTCE.
Tester/ Rater Norming and Certification	Testers will be certified by ACTFL trainers as BAT tester/raters upon successful completion of all of the requirements for certification. All certified testers/raters will be required to participate in norming activities and to recertify on a regular basis.
Score Reporting	All test scores and recorded tests will forwarded to PLTCE. PLTCE will serve as the liaison between ACTFL and the nations during the live administration of the BAT2. Speech samples will remain the property of PLTCE, and may be used for training purposes.

Annex C

NATO BILC
Benchmark Advisory Test 2 (BAT2)
Reading Test Specifications

Purpose of the BAT2 Tests	The 2nd version of the Benchmark Advisory Test (BAT2) is used by Bureau for International Language Coordination (BILC) member nations in STANAG 6001-based test norming and calibration studies. Its use as a benchmark (external measure), the results of which can be compared and contrasted with the results of national tests in listening, speaking, reading, and writing, is advisory only in nature. BILC stakeholders can use data derived from comparing 21 unique national tests with the BAT2 to gauge the effectiveness of the community's standardization and norming efforts (e.g., LTS, ALTS, and various BILC-sponsored events). Likewise, individual STANAG 6001 national testing teams can use results to compare rating consistency with other national testing teams.
Construct Definition	One of the definitions of proficiency relates to a general type of knowledge or competence in the use of a language, regardless of how, where or under what circumstances it has been acquired. Proficiency is conceptualized here as a global construct that transfers across contexts, tasks, and events, while proficiency tests attempt to sample the underlying competence by eliciting behaviors on tests that generalize to domains of interest. Proficiency can also be defined as the language knowledge that is needed to function in a future situation. The performance elicited in a proficiency test is usually measured or judged against a set of criteria, represented in a rating scale. The BAT2 will measure proficiency in reading comprehension in accordance with STANAG 6001. Reading comprehension will be measured in accordance with STANAG 6001 as the ability to: • process written discourse within a reasonable amount of time • understand the linguistic and cultural information that is unequivocally included in a spoken passage • make any author-intended inferences that are unambiguously implied by the content of the text.
Delivery of BAT2 Listening Test	Computer-based (online) delivery. This test will be administered and scored online using a tailored software product.

Test Format	The reading comprehension portion of the exam is adaptive and covers Levels 1 through 3. The exam consists of testlets, a group of five items all at the same level, which are administered one at a time. Each testlet is constructed so that it has the same difficulty of the other testlets at its level. The testlets function as a mini-test, and examinee performance on each testlet will result in one of three outcomes: 1) the examinee was able to sustain reading performance at the level; 2) the examinee was unable to sustain performance; or 3) more information is needed. Depending on examinee performance, the examinee will be given another testlet from the same level, a testlet from the higher level, or a testlet from the lower level. Once enough information has been gathered to assign a floor rating (the level where the examinee has demonstrated sustained performance) and a ceiling rating (the level where patterns of breakdown emerge) the test will end. The minimum number of testlets in a given exam is two (or 10 items) and the maximum would be seven (or 35 items) with the length of the test ranging from 10 to 95 minutes. The exam items are multiple-choice and timed by level (Level 1: 60 seconds; Level 2: 120 seconds; and Level 3: 270 seconds). Each item consists of a brief orientation, a stem, and five options (or choices). The fifth option is always "I don't know" and is included so examinees are not forced to guess when the item is above their ability level. All test materials, examples, and instructions appear in English. *An example multiple-choice exam item:* A message at the office – (representing the orientation) John, Betty called today at 12:15. She said you have a piece of certified mail to pick up. The mail room closes at 3 o'clock today. Thank you, N.F. This note tells John to: A) close the mail room at three. **B) go to get some mail. (key)** C) mail a letter for Betty. D) pick up Betty at the mail room. E) I don't know. After submitting each item for scoring, the examinee cannot return and change the item answer. If the examinee takes longer than the time allotted for a specific item, the option "I don't know" will be automatically selected and the test will proceed to the next item.

Topical Content	Content areas for all listening items are targeted to the general listener. Topics at all levels may include texts that are relevant to NATO operations. Lower level content areas will include everyday survival and work-related topics. At higher levels, content areas may include: • military and security issues • economic and political matters • scientific and technical issues • cultural and social issues.
Texts	Criteria used in text selection are the comprehension tasks and content areas described in the STANAG 6001 proficiency level descriptors. Texts are selected from a variety of authentic sources intended for the general reader of international English used in the NATO countries and NATO environment. Passages selected for this test should be self-standing and fully representative of the target level.
Types of Texts	Level 1 Personal notes, simple messages, bulletin board information, travel brochures, announcements of public events, simple descriptions of people and things, classified and other advertisements. These texts are not linguistically complex and deal with common, everyday situations requiring reading skill. The texts generally contain very basic vocabulary. Level 2 News articles about routine occurrences, magazine stories, extended biographical information, social notices, routine personal or business correspondence, and simple technical articles with detailed descriptions, narratives, detailed instructions written for the general reader. These texts deal with factual occurrences in the everyday world. They contain concrete factual vocabulary and may include linguistically complex structures. These texts have an organization that is predictable for the target language. Level 3 Editorials, commentaries, biographies with critical interpretations, criticisms, reports intended for the general reader on complex issues or specialized topics, argumentation, opinion pieces, and political analysis. These texts demonstrate a wide variety of discourse structures and a wide range of vocabulary. They contain complex argumentation, including hypothesis, supported opinion, analysis, implications and some nuances. Main ideas are often not stated explicitly and require “reading between the lines.”

Tasks and Accuracy	Reading tasks and accuracy requirements are in accordance with NATO STANAG 6001.
Test Administration	Before examination, the identity of each examinee is checked and verified. On-site proctors assist each examinee in logging onto the examination website and supervise the examination environment. All test instructions appear in English. Examinees are instructed to: • give electronic devices such as mobile phones, cameras, smartwatches, and other items to the proctor for the duration of the test; • look only at their computer screen and not allow others to see their screen; • report any suspicious activities; • avoid talking with others; • and refrain from consulting outside resources, such as dictionaries or web pages. In addition, examinees are instructed not to disclose the contents of the test to anyone, including, but not limited to, teachers, employers, or friends. While each item is timed, the examinee may move to the next item before the time has expired.
Score Reporting	Bi-level scores are reported to the POC for the country in the floor-ceiling format. The floor rating is reported on a scale of 0 to 3 which indicates the sustained ability level. The ceiling is reported as one of three levels: Random, Emerging, or Developing and indicates the level of performance at the next level.
Validation	The initial item validation work and setting of cut scores was accomplished using a modified Angoff rating procedure and Rasch analysis.
END	

Annex D

NATO BILC
Benchmark Advisory Test 2 (BAT2)
Writing Test Specifications

Purpose of the BAT2 Writing Tests	The purpose of the BAT2 writing test is to provide nations with an external criterion for validation of their general proficiency writing tests based on NATO STANAG 6001. This test is designed for NATO and partner nation military and civilian personnel who are non-native speakers of English. The BAT2 writing test assesses general English language writing proficiency up to STANAG 6001 Level 3, regardless of how it was acquired.
Construct Definition	One useful definition of proficiency focuses on general competence in the use of a language, regardless of the specific circumstances in which that language was acquired. That is, proficiency is not related to a particular curriculum, training course, set of materials, or institution. Based on this definition, a proficiency test attempts to provide a sufficiently large and varied sample of language tasks to demonstrate what examinees are able to do in that language. The aim is to measure an underlying competence which can then be generalized to similar domains of interest. The BAT2 writing test is a criterion-referenced test that will measure writing proficiency in accordance with the STANAG 6001 writing scale. That scale, with its descriptors of the tasks, content, text type, and accuracy required for each writing level, will provide the criteria for rating examinees.
Definition of writing	Writing proficiency is defined as the automatic use of one's internalized language and culture expectancy system to efficiently and purposefully communicate written language in a variety of unrehearsed tasks according to STANAG 6001 writing level descriptors for levels 1, 2 and 3.
Administration and rating of the BAT2 Writing Test	The BAT2 writing test will be scheduled online and administered under a contract with the American Council on the Teaching of Foreign Languages (ACTFL). An ACTFL approved invigilator/TCO will monitor each examination in order to obtain a ratable sample of writing proficiency. The sample will be saved using a system that is stored on a secure internet database. The digitized writing sample of each test will be independently rated by two trained and certified raters In cases of a discrepancy between the first two ratings, a third trained rater will rate the scripts independently. A final rating will be assigned when two ratings agree.

Elicitation technique overview	The BAT2 writing test will provide prompts to elicit written responses dealing with a variety of practical, social, and professional topics in formal and informal contexts as outlined in the STANAG 6001 writing descriptors for levels 1 through 3. For testing efficiency and to enable candidates to demonstrate their best language, two of the three prompts will be spiraling prompts, i.e., single topic prompts with tasks at two distinct levels. The spiraling prompts will be Levels 1 / 2, and Levels 2 / 3. The Levels 2/3 spiraling prompt will be NATO- or military- related. The third prompt will be a single prompt with only Level 3 tasks. The writing test will measure how well an examinee can write independently in English without access to editing tools under time constraints. Examinees will have access to scrap paper which will be collected and destroyed at the end of the test. The prompts are written in English; the examinee will write all responses in English.
General prompt information	The whole test time allotment and number of texts required will be included in the instructions and in the test-taker familiarization guide. Each prompt will state the suggested length of the examinee's response (i.e., several sentences, multiple paragraphs, etc.) in addition to a recommended word count range. Prompts will also have a recommended time allotment for the prompt. Each prompt will give information about the intended audience for the text (i.e., a close friend, a professional colleague, newsletter readers, etc.) and the criteria for evaluation (i.e., organizational coherence might be included for a Level 2 task, but not for a Level 1 task).
General topic information	Topics covered in the writing test will be in accordance with the STANAG 6001 descriptors. They will focus on the general language user at each level. There will be three distinct topics among the prompts for each writing test. Writing prompts will not be job specific. The Level 2/3 spiraling prompt will be NATO- or military- related. See individual levels for lists of appropriate topics.
Test Length	Examinees will have time to log into computers and input their administrative details before the timer for the writing test begins. Oral instructions will be given before the writing test begins. Once the test-taker presses the "start" button, he/she will have 120 minutes to read the prompts and write the three text responses.
Test Level 1 Content, tasks and accuracy	• The examinee needs to demonstrate the ability to write short notes, short personal letters, telephone messages, invitations, and similar texts requiring short, simple sentences but not requiring well-organized or cohesive paragraphs. • It is expected that the Level 1 writer's output can be understood by native readers used to non-natives' attempts to write. • Level 1 – Writer conveys basic information about basic personal needs in any of the following genres: basic personal letter, invitation, post card, phone message, short note.

Test Level 2 Content, tasks and accuracy	• The examinee needs to demonstrate the ability to write texts that state facts; give instructions; describe people, places, and things; and narrate current, past, and future events in complete but simple paragraphs. These texts may include simple personal and routine workplace correspondence as well as memoranda and brief reports. • It is expected that the Level 2 writer's output can be understood by native readers not used to reading material written by non-natives. • Level 2 – Writer conveys personal information about everyday, routine personal and workplace needs in one of the following genres: private letter, memoranda, brief report
Test Level 3 Content, tasks and accuracy	• The examinee needs to demonstrate the ability to write essay-length arguments, analysis, hypothesis, as well as extended explanation, narration, and description. These texts will include both formal and informal correspondence and documents for practical, social, and professional purposes. The examinee can write about both concrete and abstract topics. • It is expected that the Level 3 writer's output will rarely disturb the native reader, although there may be occasional errors. • Level 3 – Writes effective personal and professional correspondence and documents; conveys abstract concepts when writing about complex topics. Writing samples are discourse-length and must include one of the following: supported opinion, hypothesizing and speculating, argumentation or analysis
Ratable Sample	A sample will be considered "ratable" when the sample provides texts that address the prompts, and conditions and requirements have been met according to test instructions and according to the STANAG 6001 descriptors.
Rating Protocol	The examinee's performance will be rated holistically by using the STANAG 6001 writing descriptors. The rating will be based solely on the linguistic evidence demonstrated during the test. Scores will be reported as: 0+, 1, 1+, 2, 2+ and 3. Each of two raters will evaluate the writing sample independently without any knowledge of any other rater's decision. Raters requested to give a third rating will not be informed that this is a rating intended to resolve a discrepancy. Raters will submit their ratings to ACTFL for determination of a final score and for formal notification through approved channels. Neither the score nor any other information about the test will be discussed directly with the examinee. ACTFL will assign only raters who are BAT certified and have demonstrated the ability to consistently rate writing samples with a high degree of reliability.

Base Levels and Plus Levels	Plus ratings will be awarded in a non-compensatory manner; i.e., a rating of 2+ will indicate that the writing sample meets ALL of the criteria of STANAG 6001 Writing Level 2 but performance at the next higher level is not sustained, only developing. Plus levels will not be awarded for random or emerging skills at the next higher level. See table below for detail.

Level	Content *What a person can write about* *(subject areas, topic addressed; settings)*	Tasks *What a person can do with written language (tasks accomplished, attitudes expressed, tones conveyed)*	Accuracy *How comprehensible / correct the written message is* *(who can understand the writing)*	Text produced *(length and organization of texts; discourse types)*
0+	Immediate survival needs such as numbers, dates, own name, nationality, address, set expressions.	Can produce symbols; some of the most common characters. Can write only memorized material.	Spelling and representation of symbols may be incorrect. Understandable with difficulty even to a native reader used to dealing with the writing of non-natives.	Memorized words and short phrases; lists of common items.
1	Immediate personal needs (food, lodging, transportation, shopping, personal background and interests).	Can convey basic intention by writing short notes and personal letters, post cards, phone messages, invitations. Can fill out forms and applications.	Can be understood by native readers used to non-natives' attempts to write.	Discrete sentences; loose connection of sentences joined by common linking words.
1+	Basic personal needs and simple workplace-related matters.	Can readily write simple personal and routine workplace documents. Shows inconsistent and unreliable ability to write instructions; descriptions of people, places, and things; narrations of activities and short, factual accounts.	Comprehensible to a native reader used to material written by non-natives; others may need to confirm meaning with the writer.	Limited ability to connect a group of sentences coherently.
2	Everyday personal topics such as own background, family, interests, work, travel, and current events and routine topics related to the workplace.	Can write simple personal and routine workplace correspondence and related documents such as memoranda, brief reports, private letters. Can state facts; give instructions; describe people, places, and things.	Prose can be understood by a native not used to reading material written by non-natives. Individual writes in a way that is generally appropriate for the occasion although command of the written language is not always firm.	Connected prose and complete, but simple, paragraphs that contrast with and connect to other paragraphs.
2+	Practical, social, everyday professional topics, particular interests, special fields of competence and to some extent abstract topics.	Can write relatively coherent personal and informational correspondence. Can organize and elaborate on ideas in special fields of competence. Writes less effectively when supporting opinion, writing about abstract concepts, clarifying points, answering objections.	Prose can be readily understood by a native not used to reading material written by non-natives. Some errors may interfere with efforts to sustain essay-length argumentation.	Some ability to arrange a series of paragraphs into essay-length documents.
3	Practical, social, professional and abstract topics, particular interests, special fields of competence, and complex topics which may include economics, culture, science, and technology.	Can write effective formal and informal correspondence and documents. Can use language to write essay-length argumentation, analysis, hypothesis. Can convey abstract concepts when writing about complex topics.	Errors are occasional, do not interfere with comprehension, and rarely disturb the native reader.	Extended, essay-length texts.

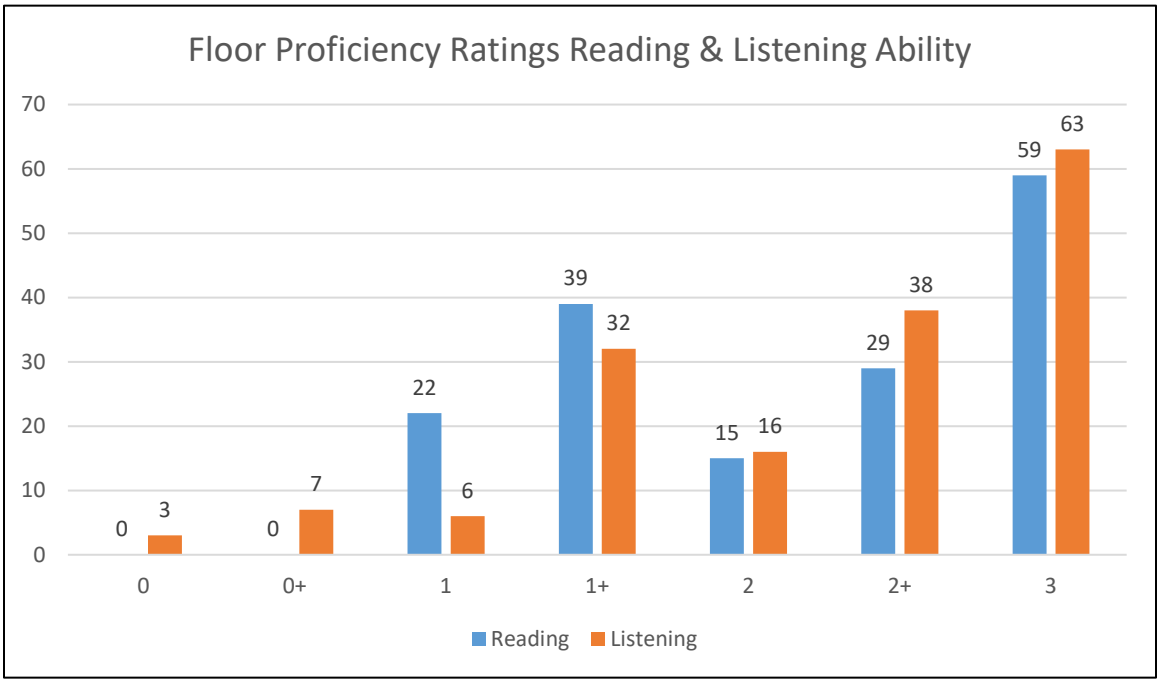
Quality Control of Rating Reliability	ACTFL will manage all quality control procedures. On an on-going (quarterly) basis, twenty percent of all writing tests will be randomly selected for quality control of ratings. All reports on results and any other information of interest will be provided to PLTCE. If PLTCE determines it is necessary to change the percentage of tests reviewed or the frequency of reporting, this change will be made.
Raters and Rater Certification	Writing Rater refers to the person who rates the test. Raters may be native speakers or non-natives whose own speaking level meets or exceeds Level 3. All testers and raters must be trained and certified by ACTFL or other organizations designated by PLTCE. Tester/Rater Norming and Certification The training will combine face-to-face instruction and online practice. Raters will be required to demonstrate the ability to consistently rate scripts with a high degree of accuracy and inter-rater reliability. Testers will be certified as BAT2 raters upon successful completion of all of the requirements for certification. All certified raters will be required to participate in norming activities and to recertify on a regular basis. An important aspect of the tester training will be the interpretation and understanding of the NATO STANAG 6001 scale and the application of this standard to operational testing.
END	

Annex E

NATO Listening & Reading Benchmark Advisory Test, Version 2 (BAT2)

EXECUTIVE SUMMARY

The following slides include the summative ratings, in aggregate, of those who took the BAT Reading and Listening tests, followed by an overview of the test development. Further, this report will provide an analysis of the test items and testlets based on data from NATO examinees, ending with a short discussion of the analysis.



BAT Listening/Reading Test Description

- The exam is designed for NATO military and civilian personnel who are non-native speakers of English. The BAT assesses English language proficiency regardless of how it was acquired. For this reason, the BAT is not related to any curriculum or language program.
- The purpose of this exam is to provide nations with an external benchmark of their general proficiency tests based on NATO STANAG 6001 (Ed. 3).
- Therefore, the scores examinees get on this test are strictly advisory in nature and not meant to replace those awarded in their respective countries, which are the official scores

BAT1 Test Development Process

- Developed by testing specialists from NATO countries.
- Test material designed, written, reviewed, and revised
    - through online interaction and
    - face-to-face meetings
- Occurred across a three-year period.
- Standard setting and establishment of cut-off scores for each level were also done through a combination of online and face-to-face activities.

BAT2 Test Development Process

- The latest update to this this exam used the existing item specifications of BAT, but retired underperforming, overexposed and dated material with approximately 50% new content.
- New items contain vocabulary, content, and listening passages with accents from the US, UK, Canada, Australia and New Zealand.
- These new items were developed by subject matter and assessment experts.
- All items were calibrated using the responses of a minimum of 100 adult ESL learners.

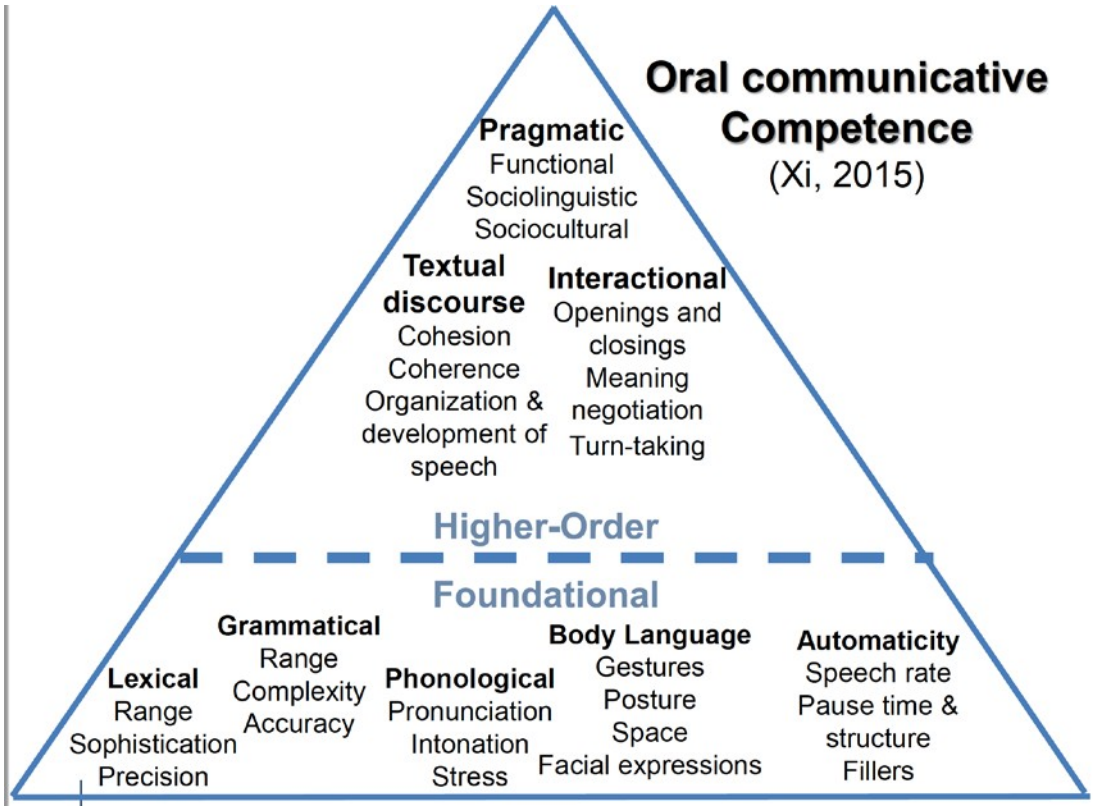
Software Delivery System

- Software allows each item to better reflect the STANAG criteria and sound psychometric processes
- Being able to go back and change responses can result in a violation of local independence

(that is-one question can give the answer to another)

- Allows timing to be added to each question
- Automaticity in processing language is part of professional and working language proficiency.
- Adaptive delivery possible—instead of 60 questions over the course of 2 hours (BAT 1) for each skill, it can be between 10 and 35 items.
- Reduces item exposure

Why Include Timing?



- Automaticity is foundational to language proficiency
- Signals different cognitive process as it shifts from deliberative thought to unconscious process
- Proficiency scale shifts from interactions with sympathetic interlocuters (Level 1) to nonsympathetic (Level 2)

Timing Parameters for Items

	Reading (Word Count)		Time Allowed		Listening (Seconds)		Number of times played
	Avg.	Max.	Seconds		Avg.	Max.	
Level 1	50	60	60		20	30	2
Level 2	150	180	120		40	60	1
Level 3	300	400	240		80	120	1

- Adaptive with items ranging from Levels 1 through 3.
- All items are multiple choice, with a single question for each text.
- All items are automatically scored either correct or incorrect by the program.
- The exam consists of testlets, a group of five items all at the same level, that are administered one at a time.
- Each testlet is constructed so that it has the same difficulty of the other testlets at its level.

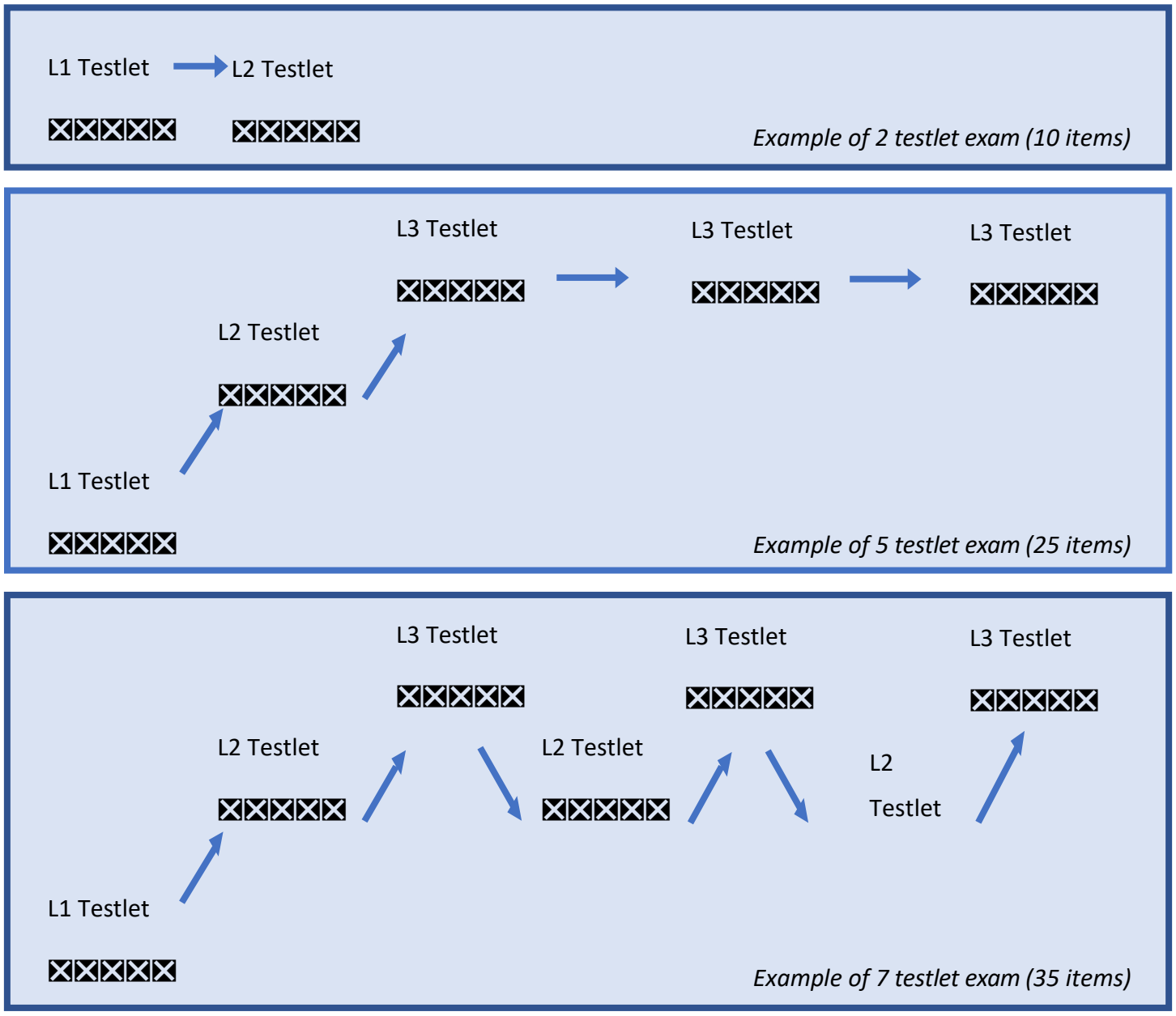
Testlets

- Functions as a 5-item, level specific mini-test
- Examinee performance on each testlet result in one of three outcomes:
    - 1) the examinee was able to sustain reading performance at the level;
    - 2) the examinee was unable to sustain performance; or
    - 3) more information is needed.
- Depending on examinee performance, the examinee will be given another testlet from the same level, a testlet from the higher level, or a testlet from the lower level.

• The minimum number of testlets in a given exam is two (or 10 items) and the maximum would be seven (or 35 items) with the length of the test ranging from 10 to 95 minutes.

Testlet Adaptive Delivery

Over 800 different decision nodes based on a pattern of strengths/weaknesses used to determine rating



Items

- Each item consists of
    - a brief orientation,
    - a stem, and
    - five options (or choices).
    - The fifth option is always "I don't know" and is included so examinees are not forced to guess when the item is above their ability level.
- All test materials, examples, and instructions appear in English.

Why "I don't know" is an option

- Psychometric model used assumes students are not forced to guess
- Paper/pencil allows students to skip and leave blank
- Computer delivered independent requires some way of verifying that the items are not left blank unintentionally before examinees move on

Testlet Creation

- Items went through cultural sensitivity review of adult ESL learners
- Each item administered to a minimum of 100 examinees
- Item difficulty parameter was calculated
- Items that did not align with intended difficulty were rejected/revised
- Testlets were created selecting items of various topics and within the major level range so that the testlets were interchangeable in terms of item difficulty

Sample Size Required for Rasch

Item Calibrations or person measures stable within	*Confidence*	*Minimum sample size range (best to poor targeting)*	*Size for most purposes*
± 1 logit	95%	16 † -- 36	30 (minimum for dichotomies)
± 1 logit	99%	27 † -- 61	50 (minimum for polytomies)
± ½ logit	95%	64 -- 144	100
± ½ logit	99%	108 -- 243	150
Definitive or High Stakes	99%+ (Items)	250 -- 20*test length	250
Adverse Circumstances	Robust	450 upwards	500

Floor/Ceiling Scoring

- Information gathered from Floor Performance is used to assign base levels (0, 1, 2, & 3)
- Information from Ceiling Performance is used to
    - assign plus levels (0+, 1+, 2+)
    - Provide feedback on development of next level criteria
        - Random (Less than chance)
        - Emerging (Greater than chance in the process of emergence)

Content by Level

- Level 1 texts
  - relate to common, everyday situations requiring survival reading/listening skills to understand very basic information. These items may include simple descriptions and narratives.

- Level 2 texts
  - relate to concrete, factual situations requiring reading/listening skills to understand detailed instructions, detailed descriptions, and narratives. Some texts may be work-related, including detailed instructions and memoranda.

- Level 3 texts
  - present complex issues, abstract language, and specialized topics requiring reading skills to understand argumentation, supported opinion, analysis, and hypothesis. Some texts may include implications and nuances. Some texts may be work-related, including technical reports and position pieces.

Receptive Test Content

Reading

- The content of reading items comes from articles written for the general reader in English-speaking countries.
- There are topics relating to everyday life and work situations, as well as higher-level texts on subjects, such as military and security issues, economics, science, and culture.
- There are also texts relevant to NATO operations.

Listening

- Audio texts come from a variety of authentic sources intended for the general listener of international English
- Audio texts may include monologs, dialogs, debates, and interviews.

Reliability

- Rasch Analysis estimates reliability using the person separation reliability index, which indicates the extent to which the scores on a given administration of the test are replicable.
- This estimate is analogous to Cronbach's alpha where the closer the value is to 1.0, the more reliable the results are estimated to be.
- Estimates of 0.70 –0.79 are considered acceptable, and estimates of 0.80 and above are good.
- Restricted range of participants will have lower reliabilities than a full range of participants

Reading Test (Reliability = .75)

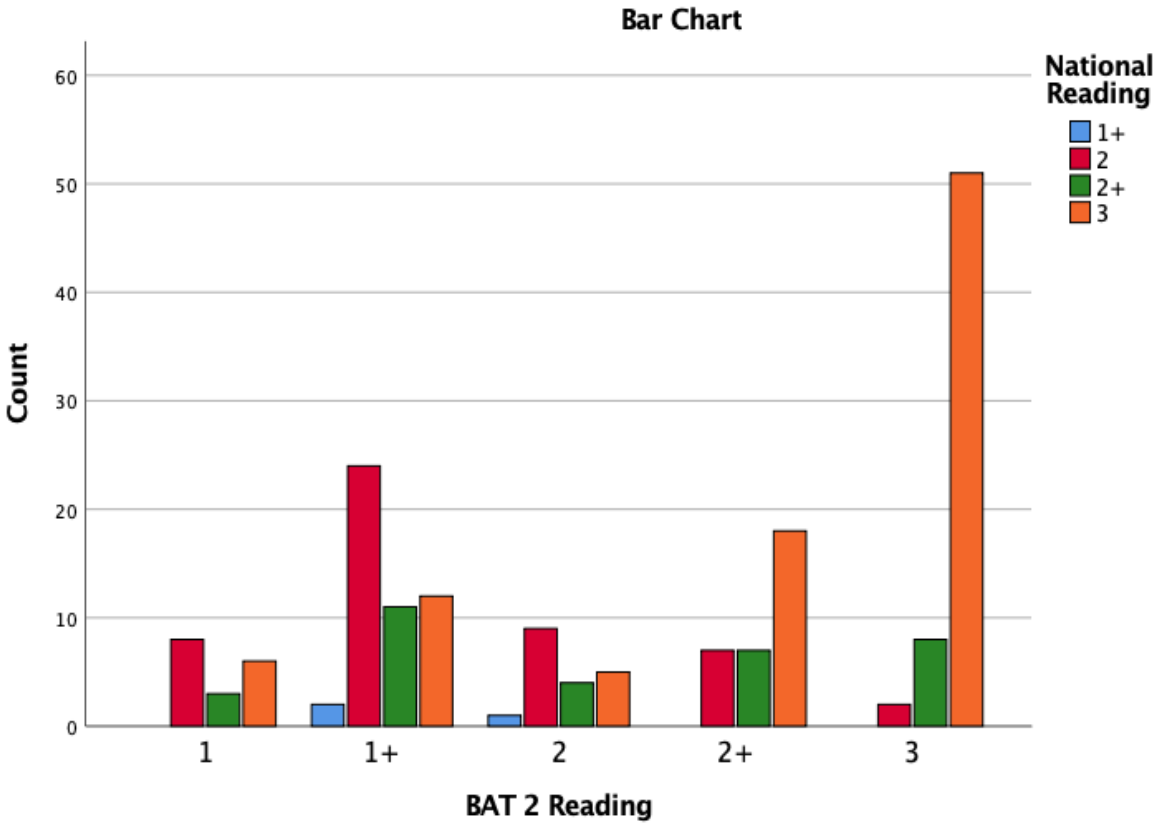
- 55 items analyzed
- 4 underfit the model (Outfit MnSq > 1.5)
- Level 1 Items
    - All difficulties functioned as intended
- Level 2 Items
    - 3 were easier than intended
    - 4 were harder than intended
- Level 3 Items
    - 7 were easier than intended
    - One testlet (3D) performed more similarly to the level 2 testlets

Items that were easier or harder than intended were still within .5 logit error associated with a calibration size of 100 respondents

Reading Testlets

		Timing (in seconds)			Exposure	
Testlet	Difficulty	Min	Max	Mean	Number	Rate
1-A	-2.74	60	245	152.59	91	0.555
1-B	-2.30	62	266	165.72	90	0.549
1-C	-2.50	63	228	146.41	94	0.573
2-A	0.69	160	587	428.53	100	0.610
2-B	0.27	160	536	382.22	115	0.701
2-C	0.76	58	570	382.64	102	0.622
2-D	-0.54	167	518	357.54	115	0.701
3-A	1.92	520	1212	849.46	61	0.372
3-B	1.21	89	1243	870.72	65	0.396
3-C	1.70	515	1233	920.49	68	0.415
3-D	0.56	327	1281	733.70	64	0.390

Reading—BAT2 vs National Tests



		National Reading							Total
		0	0+	1	1+	2	2+	3	
BAT 2 Reading	3					2	8	51	61
	2+					7	7	18	32
	2				1	9	4	5	19
	1+				2	24	11	12	49
	1					8	3	6	17
	0+								
	0								
Total					3	50	33	92	178

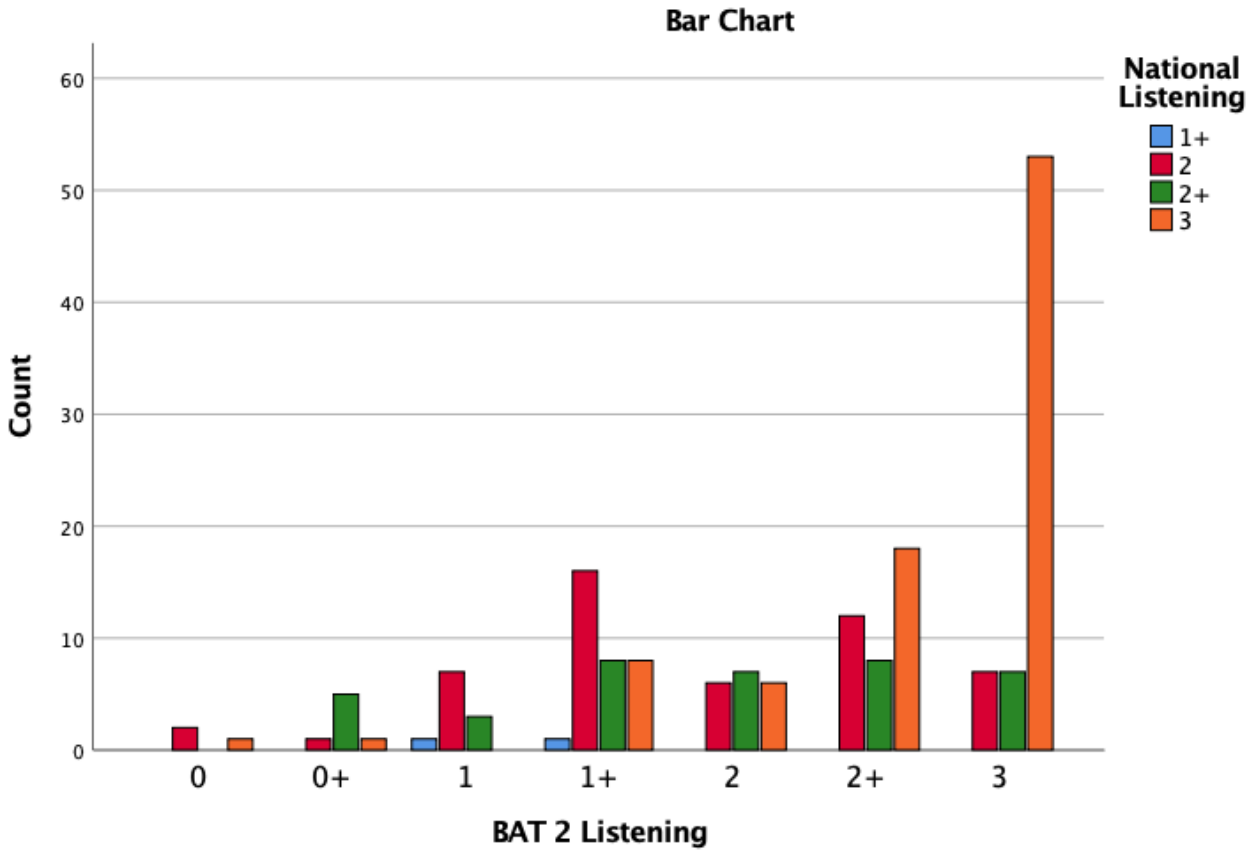
Listening Test (Reliability = .79)

- 55 items analyzed
- 1 overfit and 4 underfit (Outfit MnSq> 1.5 or <.5) the model
- Level 1 Items
    - 3 were harder than intended
- Level 2 Items
    - 4 were easier than intended
    - 4 were harder than intended
    - One testlet performed more like Level 1
- Level 3 Items
    - 4 were easier than intended
    - One testlet performed more like Level 2

Items that were easier or harder than intended were still within .5 logit error associated with a calibration size of 100 respondents EXCEPT for 1 Level 1 item.

		Timing (in seconds)			Exposure	
Testlet	Difficulty	Min	Max	Mean	Number	Rate
1-A	-2.38	71	621	218.90	140	0.848
1-B	-1.97	90	589	262.22	124	0.752
1-C	-0.64	118	764	256.43	106	0.642
2-A	-0.26	150	691	427.17	95	0.576
2-B	0.44	205	1061	441.27	86	0.521
2-C	-0.70	116	623	334.11	106	0.642
2-D	0.28	107	819	403.63	101	0.612
3-A	0.25	223	889	550.70	73	0.442
3-B	0.96	329	924	650.97	70	0.424
3-C	1.31	260	1152	672.83	82	0.497
3-D	1.61	52	804	541.27	73	0.442

Listening—BAT2 vs National Tests



		National Listening							Total
BAT 2 Listening		0	0+	1	1+	2	2+	3	
	3					7	7	53	67
	2+					12	8	18	38
	2					6	7	6	19
	1+				1	16	8	8	33
	1				1	7	3	0	11
	0+					1	5	1	7
	0					2		1	3
Total		0	0	0	2	51	38	87	178

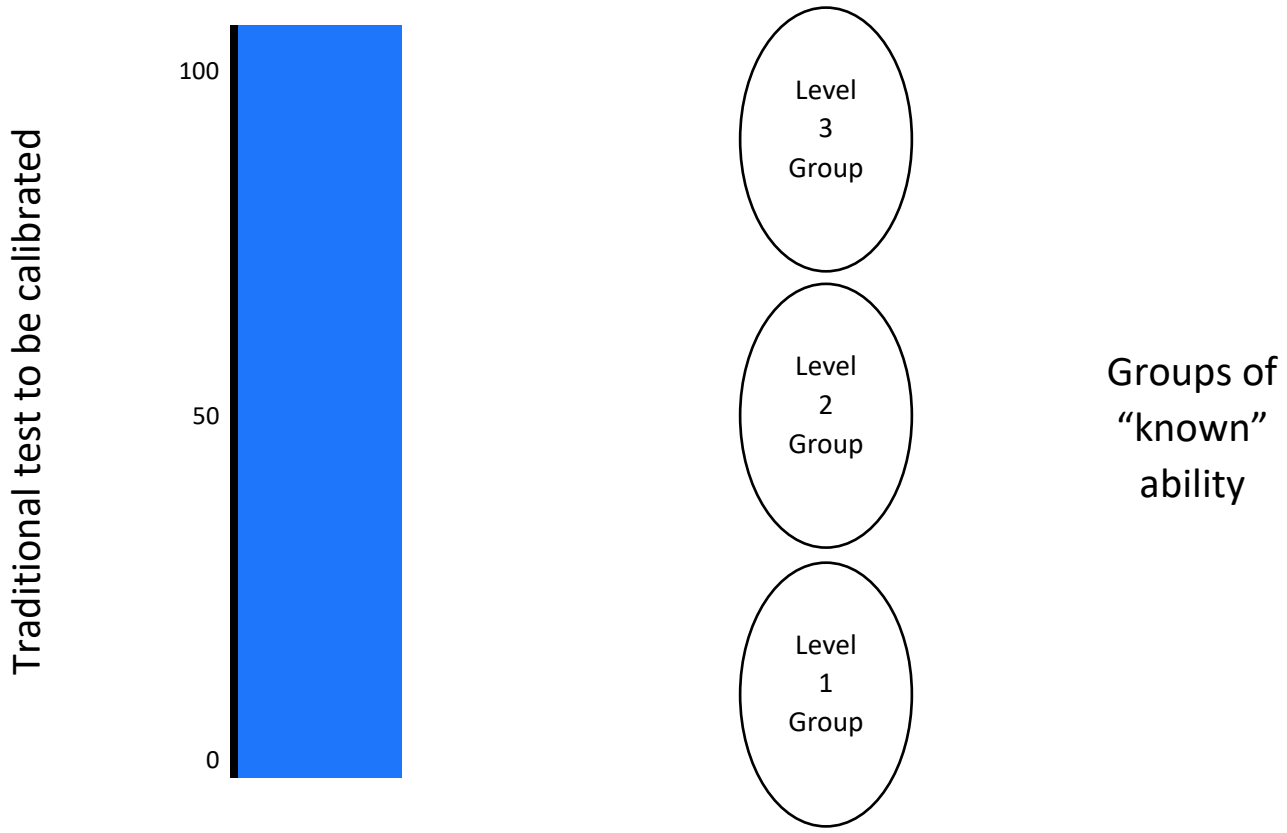
Possible Sources of Discrepant Scores

- BAT 2 was substantially shorter (16% to 58% shorter than BAT1) therefore each item has a proportionally greater weight in determining the floor.
    - More susceptible to having the test end early
- If National Tests did not contain level 1 items, the "failure" of the level 2 items might mean that the candidates were lower than 1+
- Receptive skills more subject to construct irrelevant variance such as test-wiseness (multiple choice guessing strategies, using one question to find response to another, etc.)

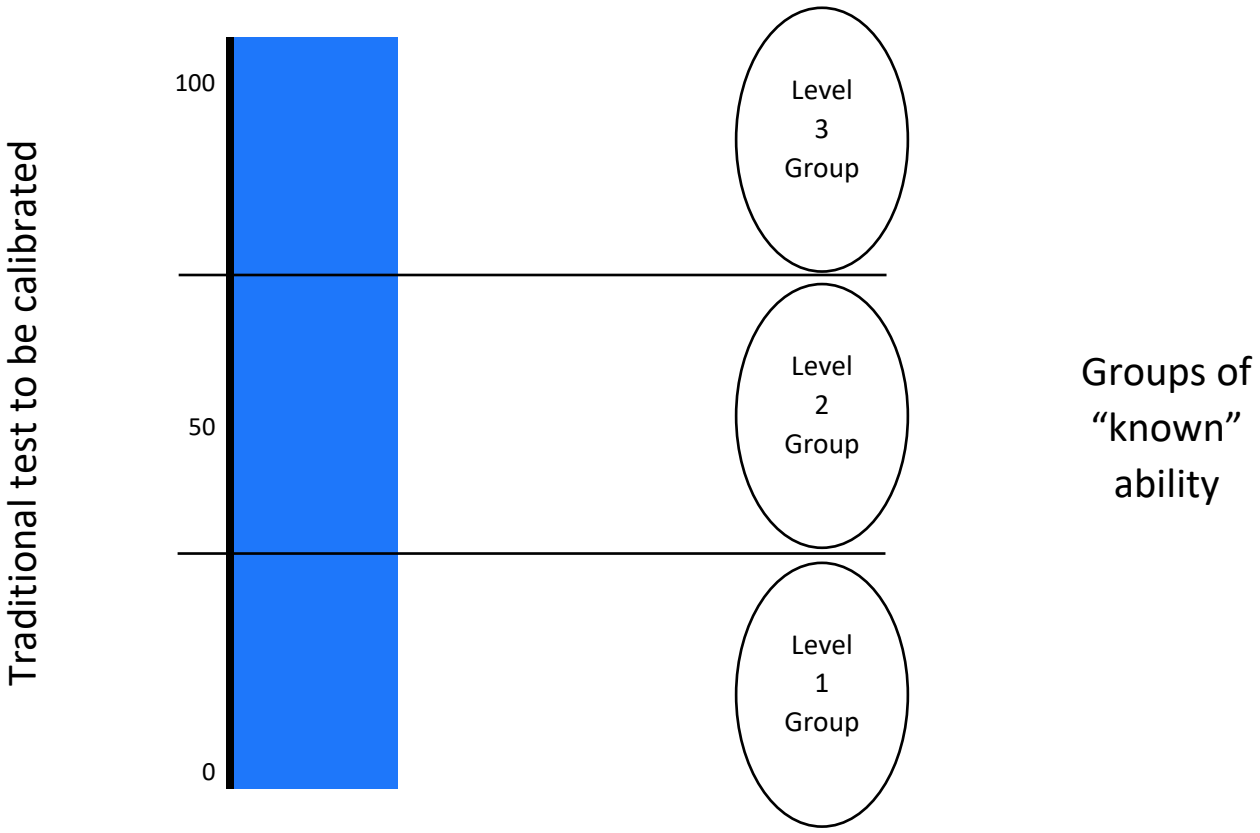
Scoring decision affect ratings

- Treating skill as ONE construct instead of THREE constructs that subsume the other constructs allows guessing to bump examinees up a level.

Traditional Method of Setting Cut Scores

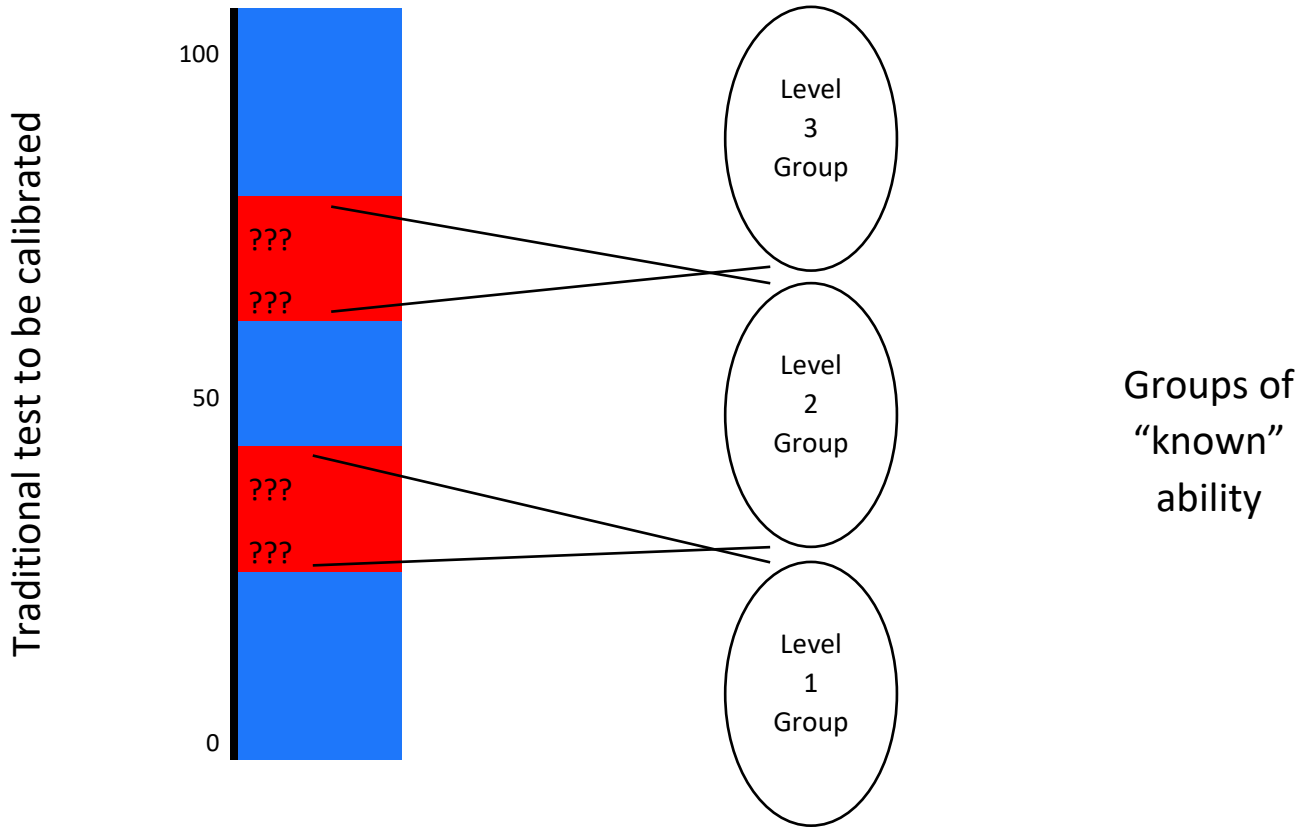


The Results One Hopes For:



The Results One Always Gets

(Some test takers score below and some score above their "known" ability)

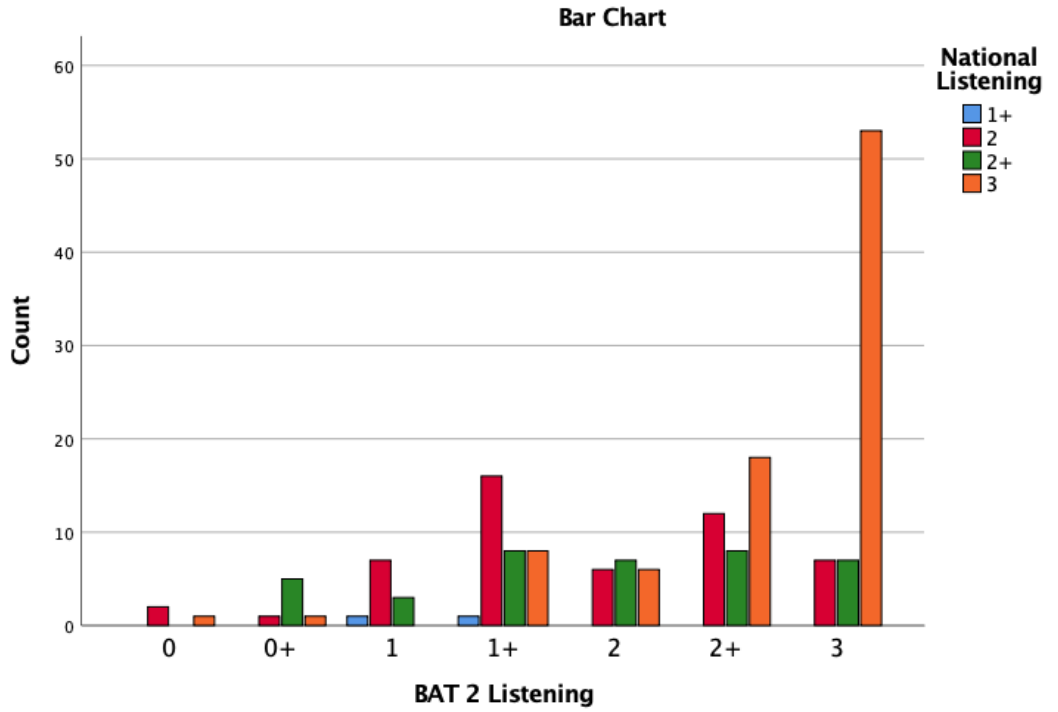


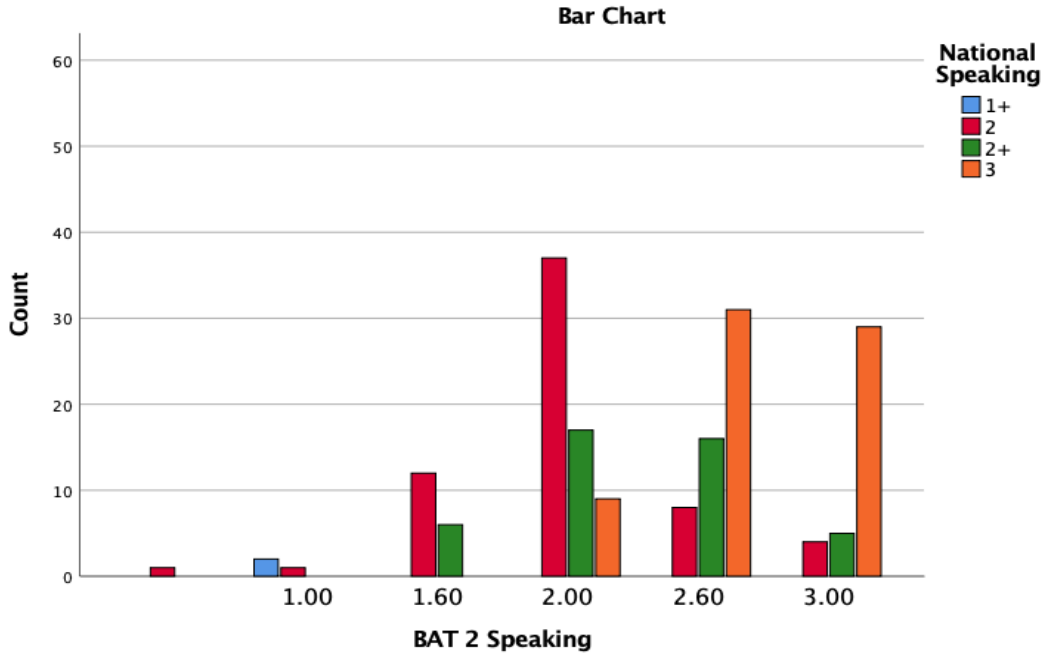
No matter where the cut scores are set, they are wrong for many test takers.

Scoring solution

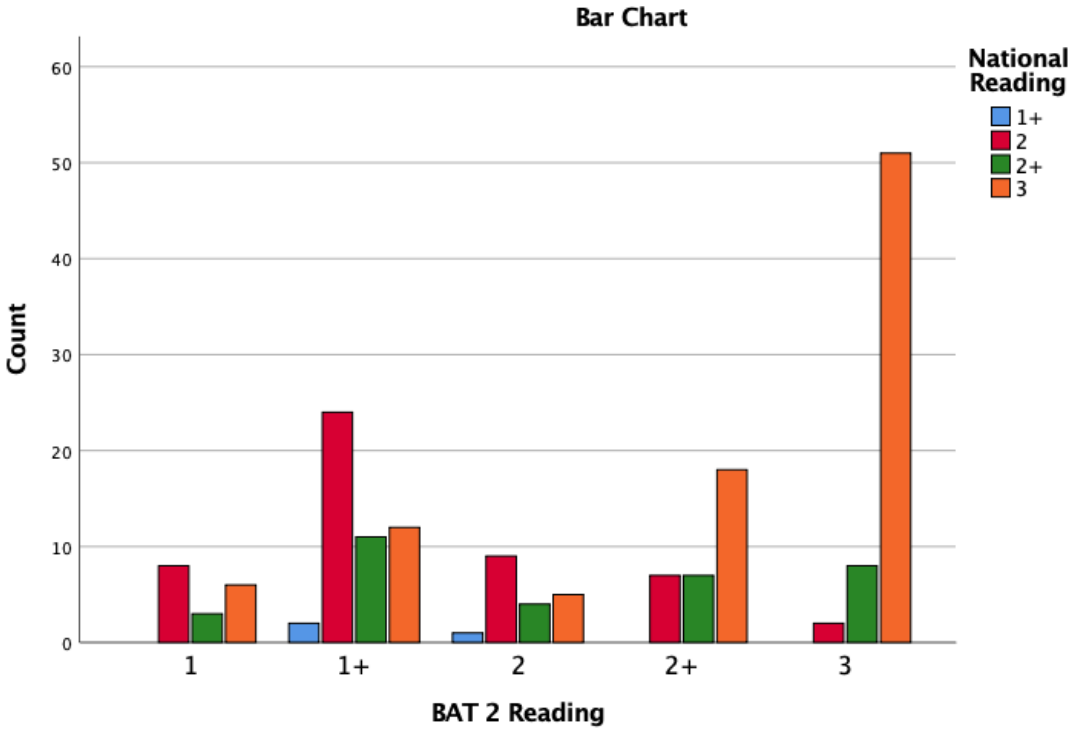
- Score each level separately!

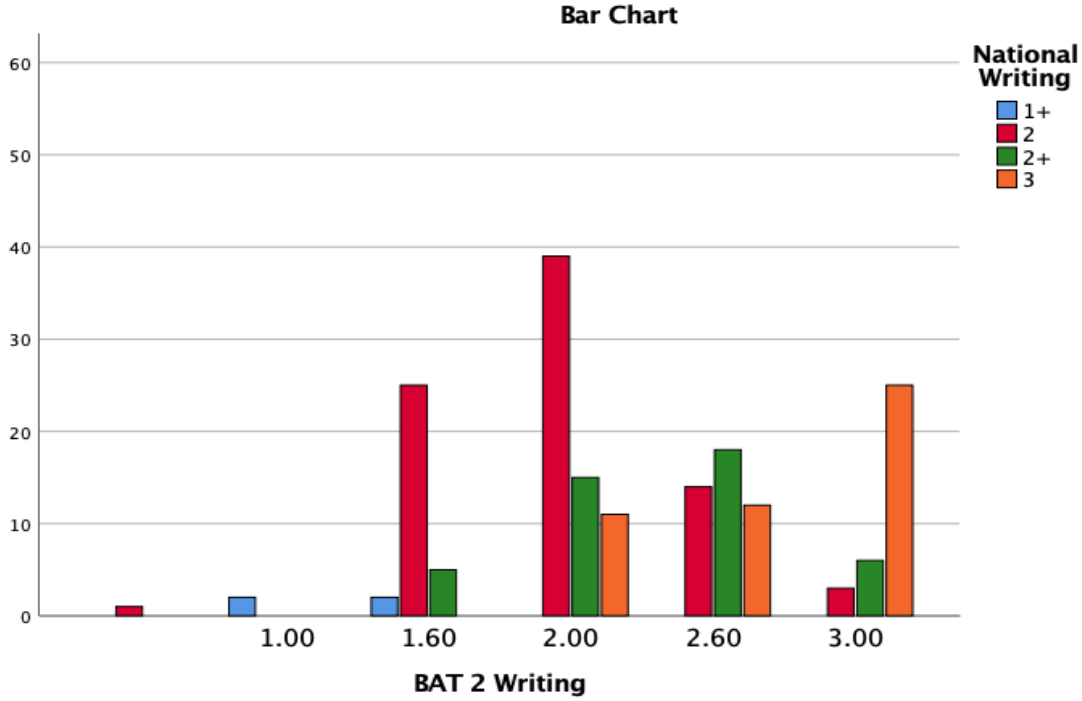
Oral Skills Comparison





Literacy Skills Comparison





Annex F

Post-BAT2 Test-Taker Questionnaire	n=103
20. The tester gave me time to answer the questions to the best of my ability.	96
19. I had enough time to show my speaking ability in English.	95
2. The BAT2 demo provided good practice for the reading and listening tests.	89
24. I understood what I needed to do to show my writing ability.	86
17. The tester asked questions that gave me the opportunity to demonstrate my true speaking ability.	81
22. I prefer to take a writing test on the computer instead of writing by hand.	76
28. I liked the BAT2 tests.	74
25. I had no technical issues while taking the writing test.	74
14. The questions, including the multiple choices options, were easy to understand.	70
3. I preferred the computer-delivered listening test to a paper and pencil listening test.	69
23. The writing test gave me the opportunity to demonstrate my true writing ability in English.	68
4. I found the listening topics interesting.	66
10. I found the reading topics interesting.	63
18. I had no technical issues during the telephone call with the tester.	59
26. I had enough time to complete the test.	53
12. The layout on the computer screen allowed easy reading of the passage and the test questions.	51
16. I felt comfortable speaking on the phone to show my ability to speak in English.	43
9. I preferred the computer-delivered reading test to a paper and pencil reading test.	41
7. Listening to the passages once was sufficient for answering the questions.	40
6. The audio quality was good.	38
13. The timing of each item was sufficient.	30
11. The BAT2 reading passages were longer than the passages on my national STANAG 6001 reading test.	-9
27. The BAT2 writing test was harder than my national STANAG 6001 writing test.	-10
8. The BAT2 listening test was harder than my national STANAG 6001 listening test.	-20
15. The BAT2 reading test was harder than my national STANAG 6001 reading test.	-22
5. The BAT2 listening passages were longer than the passages on my national STANAG 6001 listening test.	-23
21. The BAT2 speaking test was harder than my national STANAG 6001 speaking test.	-29

71 – 103 = strong agreement	51 – 70 = general agreement	1 – 50 = weak agreement	-30 – 0 = some disagreement

Discussion

Test-taker responses reflect a positive experience/impression regarding the BAT2 Speaking and Writing tests. A notable exception was the use of telephones to administer the Speaking test. Test takers were very happy with the format/protocol of the test and the performance of the testers; however, they expressed a preference for face-to-face testing. Many, also, reported problems with maintaining the connection or just being able to hear the tester clearly. A recurring concern about the Writing test was the number of tasks and the length of the test (120 minutes). On the other hand, test takers, generally, preferred taking the Writing test on the computer rather than writing it out by hand.

There was much less enthusiasm when asked about receptive-skills testing. Although test takers appreciated the relative ease of taking a computer-delivered Listening test, there was dissatisfaction with the quality of the audio passages, the number of passes, and the amount of time allowed for each item. On the Reading test, many test takers claimed that they would rather take a paper & pencil test. Most of the test takers had never experienced a Reading test with limited response time for individual items. Unfamiliarity with this feature led many test takers to express discontent with the computer-delivered test.

BAT2 test takers, generally, disagreed with statements suggesting that the BAT2 was more difficult than national STANAG 6001 tests – for all skills. This is somewhat unexpected since BAT2 scores were, on average, lower than national scores. A possible explanation is that test takers liked taking the BAT2 assessments and the outcome was not high-stakes. No one lost pay, a promotion, or an assignment based on BAT2 results.