Best Practices
in
STANAG 6001 Testing

SIGNIFICANCE AND USE

This document is intended to serve the language test developer, test provider, stakeholders, and language testing centers in their ability to provide reliable and valid tests of language proficiency IAW STANAG 6001.

Best Practices in STANAG 6001 Testing compiles research-based recommendations and principled approaches in language proficiency testing. This includes the development and uses of language tests of speaking, listening, reading, and writing for assessing language proficiency in compliance with the descriptors of STANAG 6001.

Test Purpose

The primary purpose of all STANAG 6001 language proficiency tests can be viewed as a set of three inseparable tasks:

1.) Assess an individual's unrehearsed general language proficiency level for the purpose of interoperability within NATO using the criteria captured in STANAG 6001's 0 through 5 proficiency scales, i.e. measure the individual's ability to consistently complete the real-world communication tasks in the specified situations with the level of accuracy expected in those situations.
2.) Determine which STANAG 6001 level best describes the individual's level of sustained ability.
3.) Report that proficiency level to the appropriate stakeholders.

Because the purpose of STANAG 6001 tests is to assess an individual's spontaneous abilities in frequently-occurring real-world communicative settings with the level of accuracy expected in those situation, STANAG 6001 tests are different from other tests, such as curriculum-based classroom tests and tests of technical language that would not be understood by well-educated members of that society. 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.

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. 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 namely that they must:

- be consistent with the specifications of the STANAG 6001 proficiency scales.
- test each skill modality of Listening, Speaking, Reading and Writing separately.
- treat each of the levels within each of those skill modalities as separate steps in a hierarchy of increasingly difficult and complex communication tasks.
- assess each of those levels separately, either as distinct tests or as separately scored subtests within a test battery.
- keep the items and other assessments presented at each level of proficiency aligned with the task, content/context, and accuracy triad that defines that level's specific set of expectations.
- include a representative sample of level-specific assessment items or activities at each level to justify the conclusions reached about an individual's ability at that level.
- apply criterion-referenced assessment standards and procedures to assess test-takers sustained ability at each level.

It is not possible to accurately assign STANAG 6001 proficiency scores from tests that were designed to serve other testing purposes. Attempts to extrapolate STANAG 6001 ratings from other types of tests inevitably introduce a misalignment that may overstate the test candidate's proficiency level – which in turn can lead to inappropriate assignments, operational failures, and even loss of life. For example, extrapolating STANAG 6001 ratings from any the following types of tests would yield at best an inexact approximation of an individual's true proficiency.

- Curriculum-based tests and achievement tests.
- Diagnostic tests and placement tests.
- Performance tests and prochievement tests.
- Norm-referenced tests and other tests designed to distinguish between people of varying ability levels.

Test Design

General Recommendations

STANAG 6001 tests are criterion-referenced tests.

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. The content, task and accuracy statements derived from the STANAG 6001 level descriptors form the core supporting structures of both the STANAG 6001 testing system and its associated rating system.

Each STANAG 6001 level represents a separate construct that is to be independently tested and scored.

Rating and scoring must be non-compensatory. To qualify for a rating at a certain STANAG 6001 level, test candidates must demonstrate that they meet all of the requirements for that level in a consistent and sustained fashion. For example:

- Command of a broader-than-required lexicon does not compensate for failure to accurately communicate using that vocabulary.
- On a multi-level test, correct answers of level 3 items do not count toward the total score of correct level 2 items.
- Consistent and sustained fashion thereby demonstrating mastery at that level.

Important characteristics of STANAG 6001 testing are the concepts of ratable samples and full construct representation.

- Ratable samples require that productive skills tests elicit a performance on a test that demonstrates both the 'floor' and the 'ceiling' of the language ability, and is lengthy and varied enough to allow the rater to confidently assign a score. A performance that does not clearly show a floor and a ceiling, or one that is too short or does not cover a variety of topics and tasks, is considered non-ratable.
- Single level speaking tests are effectively pass or fail and should only assign the rating of the level tested or "no rating". Plus levels should not be awarded in single level tests.
- Full construct representation. Each level test of each receptive skill should include sufficient items to test all of the level tasks in a variety of level specific content areas to the accuracy level defined by STANAG 6001.
- Good test design should outline procedures for recording the content area and task tested (with item statistics) for each item to ensure full construct representation when tests are compiled.

Test Specifications

Test specifications should exist for all STANAG 6001 tests and reflect the test design and best practices. There should be separate test specifications for each skill.

Test specifications should be easily updated, living documents.

Testing specifications for the Receptive Skills should include:

- The purpose of the test
- The target testing population
- Language of the instructions, orientations and stems (mother tongue, target language)
- Levels to be tested (single, bi-, multi-level tests)
- Total test time
- Length of reading texts in words and speaking texts in time (minutes and/or seconds) per level.
- The number of times a listening text can be heard and the number of different voices in the text
- The number of texts per level and items per texts
- Item formats (multiple choice, constructed response, other)
- Skills and subskills tested
- Topical content areas for each level
- Number of correct items for each level to demonstrate mastery
- Plus level determination, i.e., Random, Emerging, Developing and Sustained (REDS)
- Test administration details (computer-delivered vs paper and pencil; audio through headphones or not)

Recommended practices

Authentic texts (written or spoken) by native speakers for native speakers should be used for Levels 2 and 3. Level 1 texts may be authentic or semi-authentic but should be genuine[1].

In order to ensure a representative sample, each level of the reading and listening tests should have 15-20 texts with a variety of topical content and 1-2 items per text.

Reading texts should be self-standing and fully representative of the target base level. Suggested text lengths and number of items:

- Level 1= up to 60 words (1 question per text)
- Level 2= up to 150 words (1 or 2 questions per text)
- Level 3= up to 300 words (1 or 2 questions per text)

Listening texts should be should be self-standing and fully representative of the target base. Listening texts should be clear without interference or background noise. The length of the listening text should be limited to avoid testing memory. Suggested text times and number of items:

- Level 1= up to 45 seconds (1 question per text)
- Level 2= up to 60 seconds (1 or 2 questions per text)
- Level 3= up to 90 seconds (1 or 2 questions per text)

Reading and listening texts should represent a balance of the varieties of the English spoken in NATO contexts.

[1] A genuine text is one that is accepted by educated native speakers as an authentic text.

Testing specifications for the Productive Skills should include:

- The purpose of the test
- The target testing population
- Language of the instructions (mother tongue, target language)
- Levels to be tested (single, bi-, multi-level tests)
- Structure of the speaking and writing tests
- Total test time
- Number and type of tasks and accuracy statements to demonstrate mastery at each level
- Topical content areas for each level
- Test administration details (speaking tests recorded or not; handwritten or electronically-written writing tests; aids permitted or not for writing tests)

Speaking tests should be recorded. If the two raters disagree on the rating, a third rater can listen to the recording. Recorded speaking tests can also be used in norming or training sessions.

Role plays should be one of the speaking test tasks. Role plays are a proven method for eliciting conversational speech and appropriate register.

Scoring Receptive Skills Testing

All tests require answer keys.

Receptive skills tests should be double-marked for accurate scoring.

Answer keys for constructed- response questions should list all possible answers.

Clear, consistent procedures should exist for agreeing and scoring correct answers that are not on the answer key.

Rating Productive Skills Testing

Tests of productive skills should be rated by a minimum of two trained and normed raters.

Speaking tester roles should be clearly defined

- Interlocutor only
- Interlocutor and rater
- Rater only
- Reviewer

Rating scale rubrics should be produced for rating tests of the productive skills to help ensure non-compensatory scoring.

Item and Prompt Development

General recommendations

Item writers should not work on their own, but in and as a team

Before the team starts developing items, there should be consensus about what and how to test (as laid out in the test specs)

Make sure that each item and prompt measures unrehearsed general language proficiency, and not discrete-point grammar and vocabulary, or curriculum-based performance.

Quality control measures

Keep the testing section separate from the teaching section in order to avoid conflicts of interest

Item writers should be properly trained and (re)normed against the STANAG 6001 scale

Keep all testing materials secure at all times and provide access only on a need to know basis

Update items regularly to avoid them becoming outdated or compromised

Item & prompt writing

Item/prompt writers must have a thorough understanding of the C/T/A statements.

Item/prompt writers must have a thorough understanding of the test specifications and write items and prompts in accordance with specifications.

Item/prompt writers should develop items/prompts individually but moderate as a group.

Recommended practices for receptive skills testing:

Develop (depending on the experience of the test writers) approx. 40-60% more items than needed according to the specs.

Use (with the possible exception of Level 1 items) authentic texts.

Ensure representative sampling of text types (genres), topics and tasks

Select texts that are appropriate with regard to text type, length and difficulty for the level to be tested.

Text editing should be purposeful and minimal. Edited texts should be genuine (a text that is accepted by educated natives as authentic).

Keep (originals of) texts for reference and future use.

Develop listening items from the recording, not from the transcript.

Avoid testing memory instead of listening skills.

Follow established, commonly accepted item writing guidelines.

Make sure that test instructions are clear and concise.

Use an item bank to maintain the item pool.

Maintain metadata on items in item bank.

Recommended practices for productive skills testing:

Avoid tasks that are vague and open to individual interpretation as to what the task is and how much and what kind of language is to be produced.

Prompts should elicit lengthy samples of a variety of speaking/writing objectives to increase both reliability and validity.

Provide clear rubrics and rating criteria.

For speaking: develop a procedure that establishes the candidate's floor and ceiling.

Moderation Practices

Item moderation is an essential step in test development in order to:

- ensure that the text, task and level of each item are aligned.
- perform a quality control check and identify any flaws or errors in the items or prompts.
- determine if the items should be kept as is, revised or discarded.

Moderation process for listening and reading items

A board or panel of no more than 5 or 6 members should be convened to review the draft items. The members should include the head of the board, the writer(s) of the items, test development team members and external reviewers.

Appropriate BILC tools and checklists (alignment tables and checklists, item review checklist, etc.) should be utilized.

In preparation for the moderation session, items to be reviewed should be complete but without the suggested correct answer.

During the moderation session:

- panel members should try to answer each item as if they were taking the test
- panel members carefully review each draft item, using a checklist
- item writers should provide explanations and keep records (written or taped) of all suggestions and comments
- the head of the board should ensure that proper moderation procedures are followed and should make the final determination of the status of the item—to keep as is, revise or discard

After the moderation session, the item writers will take follow up action on revising or improving items.

- All notes and successive drafts and re-workings of the items should be kept. The rejected wordings and extra options may make any necessary future revising easier.

Moderating speaking and writing prompts

Speaking and writing prompts should also be moderated by a panel or as a minimum reviewed by other members of the testing team. Moderation should determine if the prompts should be kept as is, revised or discarded. When moderating prompts, special attention should be paid to whether:

- the language of the prompt models the target language level
- the task is at the appropriate level
- the task is clear
- the audience is clear
- the task will elicit the language expected

Rater Norming

Rater norming is the active engagement of a community of raters to align themselves to a set of professional standards (e.g., STANAG 6001 skill level definitions) while assessing Speaking and/or Writing (ratable) samples produced within a second-language testing context.

Benefits

- Process reinforces a specific and systematic approach to assessing test performances by focusing on tangible evidence of Speaking/Writing abilities
- Candidates receive the same rating for a Speaking/Writing test performance regardless of the individual rater (team).
- Process refines elicitation techniques and helps raters internalize the test structure/protocol

Preconditions

When assessing Speaking/Writing ability, STANAG 6001 skill level descriptions, Content/Task/Accuracy statements, and scoring rubrics are only as good as the raters using them.

Ratable samples which reflect floor and ceiling performances are essential for successful rater norming.

Facilitators for rater norming must have a thorough understanding of:

- STANAG 6001 skill level descriptions
- Content/Task/Accuracy statements
- Scoring rubric(s) for the skill(s)
- Inter- and intra-rater reliability of Individual raters
- Training expectations of the community of raters
- Speaking/Writing samples used for the norming session

Recommended Stages in Rater Norming

Facilitator thinks aloud through several benchmark samples to model the rating process

- Facilitator should model rating of samples at various levels (followed by questions)
- The goal is for raters to come to a consensus rating, not just agree with the facilitator
- Norming should allow raters to "take ownership" of the rating process

Raters independently score a few samples at various base and plus levels

- In plenary, raters determine whether a given sample is or is not ratable
- Raters analyze performances considering STANAG 6001 skill level descriptions, CTAs, and the scoring rubric
- Raters discuss samples, looking for patterns of consistent and inconsistent ratings
- Raters discuss application of scoring rubric factors to reconcile inconsistent ratings
- Raters compare group (consensus) rating with expert rating

Raters score more samples at various base and plus levels

- Samples may be more challenging with more borderline decisions
- Facilitators might encourage raters to work, at first, independently, then to discuss ratings in pairs or small groups.
- Raters, again, compare consensus ratings with expert ratings
- If raters are generally consistent, facilitator might decide to focus on specific problem areas

Facilitator wraps up the norming session by reemphasizing the importance of efficiently-elicited ratable samples and reliable, objective scoring of candidate performances

Signs of Rater Drift (Construct-Irrelevant variance)

Moral Dimension

- Severe – as reflected in Inter-rater Reliability statistics, ratings are, across the board, significantly *lower* than those of other raters.
- Lenient - as reflected in Inter-rater Reliability statistics, ratings are, across the board, significantly *higher* than those of other raters

Halo Effect – as reflected in intra- and inter-rater reliability statistics, certain ratings are significantly *higher* than the samples would warrant. For example, a rater, who is also a teacher, might consider classroom performance (not part of the test sample) in determining a final rating.

Central Tendency – as reflected in Intra- and Inter-rater Reliability (IRR) statistics, the rater awards a narrower range of scores centered on expected outcomes. For example, in testing following a STANAG Level 2 class, the majority (if not all) might receive a Level 2 rating even though there might have been higher- or lower-level performances.

Frequency of norming

Norming is recommended when there are new testers, if there is evidence of rater drift or problems with inter-rater reliability, and prior to a major testing session.

Considerations for Future Norming Sessions

- Frequency of norming or specific signals for the need
- Duration of each norming session
- Range of samples against which to be normed (e.g., lower-, mid-, upper-, or full-range)

Trialling

Trialling is an umbrella term to indicate, "*trying out test materials to gather various types of information about their performance and measurement characteristics*". As such, trialling can be divided into piloting and pretesting.

Piloting is often used to refer to a form of exploratory testing involving a small number of test-takers (among which, if possible, (near-) native speakers of the language) who can provide useful feedback on different performance aspects of the test materials. The focus is on collecting data on individual items, prompts and rubrics: their wording (intelligibility, grammatical correctness), effectiveness (eliciting the expected response), acceptability/suitability, etc. Piloting is usually done first before any pretesting is carried out, and normally takes place in an informal setting (not under testing conditions). Its main purpose is qualitative analysis, as the piloting group is usually too small to conduct any meaningful quantitative analyses.

Pretesting is administering – to a large(r), representative sample of the intended test-taking population and under official testing conditions – a test that closely resembles the final, operational test. In STANAG 6001 testing or other types of criterion-referenced testing the sample should include test-takers in a wide range of ability levels. Pretesting is typically used for receptive skills testing, as its main focus is on collecting quantitative data, both on individual items and on the test as a whole. Although the data analysis is primarily quantitative, a qualitative analysis (e.g. by using a questionnaire) may be an additional objective.

Trialling design phase

- Design the trialling procedures and include sufficient time for trialling in the test development action plan
- Allocate resources (personnel and software tools) for trialling
- Procure software tools and ensure that personnel are trained to conduct data analysis
- Determine which statistical methods and software tools to use for the data analysis

Piloting

- Identify native speakers, language teachers or professional colleagues who are willing to provide qualitative feedback on speaking or writing prompts, receptive skills test items or rubrics
- Collect and analyze their feedback and revise the prompts or items as necessary

Pretesting

- **Preparing for the pretesting**
  - Define criteria for selecting the right population for pretesting
  - Identify test population and invite them to the testing session
  - Decide on incentives for test takers
  - Determine what quantitative and qualitative feedback to collect
  - Coordinate with other testing teams if pretesting abroad
  - Prepare relevant feedback questionnaires and pilot them
  - Prepare clear test administration instructions for the proctors and test takers
  - Familiarize the test takers with the format (examples, demo, etc.)
  - Assemble moderated items into tests
  - Ensure that the pretesting conditions are similar to operational testing conditions

- **During test administration**
  - Ensure that the test takers are briefed on what to expect
  - Collect qualitative data from test proctors and test takers (via questionnaires, interviews, etc.)
  - Thank the volunteers (proctors and test takers)

- **After pretesting**
  - Officially thank the host nation if trialling was conducted abroad
  - Analyze the data both quantitatively and qualitatively in accordance with the trialling design plan
  - Decide to keep, revise, or discard pretested items
  - Document metadata for the items and include in the item bank
  - Make a list of lessons learned
  - Produce technical reports
  - Revise test specifications if necessary
  - Assemble final test versions

Test Administration

General recommendations

Administer tests uniformly to all test takers, otherwise scores will vary due to construct-irrelevant factors.

Provide stakeholders with guidance on who is eligible for testing sessions (for example, length of time after last test required before retesting; minimum screening test score, etc.)

Publish a formal schedule of testing sessions with the time, date and location of testing sessions and require advance registration.

Post familiarization guides on the MoD website or otherwise make them available to test takers and stakeholders. Familiarization guides should include, as a minimum, information on test item formats and procedures for answering test questions, sample test questions, and information on test delivery and time limits.

Anonymity of test takers should be guaranteed – e.g., use codes vice names.

Teachers should not test their own students.

Record (video or audio) speaking tests.

To prevent testing fatigue, testers should not conduct more than 8 speaking tests per day with breaks in between tests.

Equivalent forms of the test are needed to prevent test compromise and to facilitate re-testing when necessary.

Security procedures must be in place for the storing and movement of test materials.

Establish and publish clear policies on the following:

- cheating and the penalties for cheating.
- dealing with interruptions or extenuating circumstances; i.e., latecomers, examinees who become ill during the test administration, a power outage, etc.
- receiving and responding to complaints and appeals for reconsideration.

Prior to test administration

- Proctors/invigilators must be trained. They should be given the standardized test administration procedures that they should read verbatim.
- Proctors/invigilators should be given an official list of approved names prior to the testing session.
- The room where the test is to be administered should be checked prior to each testing session. This entails seeing that there is sufficient distance between test takers to prevent cheating, that the lighting and ventilation is adequate, and checking that the equipment is functioning properly (clocks, headphones or loudspeakers for listening tests, video/audio recording equipment for speaking tests.)

During test administration

- Ensure conditions are uniform for all test takers.
- Proctors/invigilators must be committed to test security and report any violations.
- Proctors/invigilators must remain in the testing room at all times during the test.
- Anonymity of test takers should be guaranteed – e.g. use codes vice names
- Test takers should be given clear instructions on how to behave during the test, what the time limits are, and what the policy is on mobile phones, electronic devices, reference materials, taking notes, cheating, etc.
- Test takers should be informed of the time limits and how to track the time.
- Test takers should be advised on how they will receive their results.
- Test takers should be familiar with policies for complaints/recourses, etc.

After test administration

- Secure tests and test papers – test material should not be left unattended.

End