BILC POLICY RECOMMENDATION
on the
Portability of STANAG 6001 Language Certifications

Introduction

As a result of the changing role of the defence organizations with regard to the employability of retiring military personnel, the portability of NATO STANAG 6001 language certifications to civilian contexts has become an important and recurring issue in many NATO nations. *STANAG 6001 Language Proficiency Levels (Ed 5v2)* is the language proficiency scale used throughout NATO, but it is not widely known or recognized outside the alliance. The language proficiency scale most commonly used in civilian educational systems in Europe is the *Common European Framework of Reference for Languages* (CEFR). As such, the prevailing opinion is that a conversion of STANAG 6001 certifications into CEFR qualifications would serve as a means of facilitating military retirees' integration into the civilian workforce. The question of portability is further related to national policies concerning recruitment and retention of national military personnel. The fact that to date there are no official correspondences between STANAG 6001 and CEFR language proficiency levels may impede the recognition of previously acquired language competencies, hinder mobility of employees and lead to double certification at a considerable cost.

A BILC Working Group, consisting of language training and testing experts from 8 NATO nations, investigated the possibilities of establishing equivalencies from STANAG 6001-based testing and certification systems to those based on the civilian standards of the CEFR. After much discussion and drawing on recent empirical and theoretical research, the Working Group came to the conclusion that there are substantive technical issues which prohibit accurate comparisons of STANAG 6001 certifications with civilian certifications and vice versa. The purpose of this paper is to discuss the main problem areas, and to provide guidance to member nations on how to deal with the issue of equating different language scales.

Discussion

The conversion of language qualifications in terms of one scale into levels of another presupposes that both language scales can be equated and aligned. At the surface level, the two frameworks appear to have several features in common: both the CEFR and STANAG 6001 are intended to serve as a common yardstick for reporting and comparing measures of general language competence; each system presents a hierarchy of language development; each divides its hierarchy into stages via prose descriptions; and both employ "can-do" statements. In addition, a number of phrases used in the respective level descriptions are similar. However, a comprehensive comparison of both frameworks will need to go beyond the alignment of similar statements in the level descriptions (Clifford, 2012: 49). It would need to take into account the theoretical bases for the schemes and the contexts in which they are used (Green, 2012: 84). In particular, the differences found in the purpose and interpretation of the scales, and in the methodology and scoring criteria used in the assessment procedures based on each of those frameworks warrant closer scrutiny.

Characteristics of the scales

The CEFR scale is primarily designed as a tool for reflection, communications and empowerment, intended mainly for language learners and teachers. The CEFR was not conceived as a harmonisation project and does not prescribe which methods to apply for language learning or testing. Rather, users are invited to adapt the CEFR illustrative descriptions to the specific context concerned (Council of Europe, 2001: 7). The STANAG 6001 scale, on the other hand, 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. (NATO Military Agency for Standardization, 1976). Despite other possible uses of the STANAG 6001 system, its predominant use has been for language assessment for high-stakes purposes.

These differences in purpose and orientation have significant implications for the ways in which both frameworks are operationalized. Given the imperatives of safety and security in a military environment, inaccuracies in assigned language proficiency levels can easily lead to operational failure, and in the worst case, to loss of life. Consequently, each proficiency level of STANAG 6001 constitutes a distinct threshold. The levels are strictly delimited by explicitly describing both 'floors' and 'ceilings' of language performance: what individuals can *and* cannot do with the language at particular levels.
In contrast, and consistent with its broader goal of promoting language learning and developing learner autonomy (Little, 2012: 71), the CEFR uses mostly positively stated "can-do" statements. In order to make the framework flexible enough to be adapted for use in various instructional and work-related settings, the boundaries between CEFR levels are purposely kept vague (Lowe, 2012: 99). Whereas the CEFR can be regarded as a multipurpose framework for learning languages and for keeping track of progress, STANAG 6001 constitutes a fixed scale for accurate measurement (ibid: 104).

Establishing a crosswalk between the two frameworks is further complicated by the fact that the STANAG 6001 scale covers a wider span of language ability (from *no proficiency* to *functionally equivalent to a highly articulate native speaker*) than the CEFR. However, this wider range is subdivided into fewer stages: five base levels (not counting Level 0) vs. six CEFR bands. Even assuming that it were possible to align levels of both frameworks, it would unavoidably result in an overlap where a certain STANAG 6001 level corresponds with at least two levels of the CEFR, and conversely, more than one CEFR level may fit in one particular STANAG 6001 level.

Assessment methodology

As a consequence of the differences in intended purpose and main use, both frameworks employ fundamentally different assessment methods.

- *STANAG 6001 tests*

  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 real-world communication tasks in 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.

  Because the purpose of STANAG 6001 tests is to assess an individual's unrehearsed, curriculum-independent abilities in frequently-occurring real-world communicative settings, STANAG 6001 tests are different from both curriculum-based achievement tests and performance (job-related) language tests. Unlike general proficiency tests, achievement tests measure only ability to perform tasks using rehearsed language that is included in a specific course of study. Performance tests are usually task-oriented and more limited in scope than proficiency tests: the language ability measured by performance tests applies only to the specific, job-related tasks being tested and cannot be transferred into general proficiency in the foreign language. Attempts to extrapolate STANAG 6001 ratings from other types of tests, such as achievement or performance tests, inevitably introduce a measurement error that may overstate the test candidate's proficiency level – which in turn can lead to inappropriate assignments and failure to accomplish critical job duties and operational tasks.

  The criticality of accurate measurement of language competence for military job requirements has placed a strong emphasis on standardization of the STANAG 6001 testing protocols. All STANAG 6001 tests 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 task, condition, and accuracy (TCA) expectations form the core-supporting structure of both the STANAG 6001 testing system and its associated rating system.

Another important characteristic of STANAG 6001 testing is that the tester is required to obtain a 'ratable sample', i.e. to elicit a performance on a test that demonstrates both the 'floor' and 'ceiling' of the language ability, and is long and varied enough to allow the rater a confident assignment of 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, is considered non-ratable.
To qualify for a rating at a particular STANAG 6001 level, test candidates must demonstrate that they meet *all* of the requirements for that level in a consistent and sustained fashion. This makes these ratings strictly non-compensatory. For instance, command of a broader-than-required lexicon does not compensate for failure to accurately communicate using that vocabulary.

- *CEFR tests*

In comparison with the STANAG 6001 framework, the CEFR is clearly more open and less prescriptive, having been designed to encompass a wider range of activities and purposes, varying from low-stakes curriculum-based achievement testing to high-stakes testing of academic language skills for university admission. The CEFR does not stipulate a particular methodology to be used in assessment, nor does it offer measures for standardizing raters or for eliminating divergent judgements. Rather, users have to determine themselves which test type and method and rating procedures best fit to their needs (North, 2014: 10).

However, precisely due to the absence of guidelines and protocols, the CEFR is not fully appropriate for high-stakes testing, as pointed out by leading CEFR experts (Trim, 2012b: xvi; Saville, 2012: 66). There are no CEFR tests in the same way as there are applications of the STANAG 6001 guidelines for the standardised assessment of language competence. Hundreds of tests and assessment procedures have been developed with reference to the CEFR, covering numerous different purposes and contexts of use. These tests are often unregulated and highly diverse in terms of their quality (Saville, 2012: 60). An empirical study on the comparability of a large number of CEFR examinations revealed that these tests were not easily comparable because they measured different language abilities, using different assessment methods based on different interpretations of the CEFR descriptors (Saville and Gutierrez Eugenio, 2016: 10). Therefore, evidence that two language tests have been related to the CEFR at the same level by no means indicates that their results are interchangeable, which in turn seriously decreases the possibilities of any meaningful comparison of language certificates based on different CEFR tests.

The features that characterize STANAG 6001 testing are not usually incorporated in CEFR tests. In contrast with STANAG 6001 tests, in CEFR tests there is normally no requirement for test takers to meet *all* of the criteria for a specific level in a consistent and sustained manner. Raters are not necessarily trained and renormed to be able to apply the scale consistently, nor is it mandatory to ascertain the 'ceiling' of the candidate's language ability. Typically, proficiency levels in a bi- or multilevel CEFR test are not tested and rated separately level-by-level, and scores are cumulative – which implies that weaker performance in one area may be compensated by stronger performance in another, resulting in higher ratings. Performances on CEFR tests would often be judged as 'non-ratable' from the perspective of the STANAG 6001 system, either because the ceiling cannot be established, or because the test measures mainly rehearsed language, or the requirements for task, condition and accuracy (TCA) are not aligned. Even when there is an alignment of the TCA elements, different scoring procedures will produce divergent test results.

Whereas a STANAG 6001 test is by definition a high-stakes proficiency test assessing language competence of adults working in a military environment, CEFR tests may have very diverse purposes, ranging from a general proficiency test to a curriculum-based achievement test for young learners, a low-stakes placement test or a job-related performance test. As noted above, there are significant differences between the various types of language test. Empirical studies show that it is not possible to accurately assign STANAG 6001 proficiency scores from tests that were designed to serve other testing purposes (Clifford, 2001). The fact that it is not usually possible to determine from a CEFR language certificate the type of test on which the rating was based greatly undermines the value of these certificates.

Quality assurance mechanisms

The highly standardized approach to testing in accordance with STANAG 6001 requires a regulating body that monitors the common interpretation of the level descriptors and consistency in the assessment methodology throughout NATO. The custodian for the STANAG 6001 framework is the *Bureau for International Language Coordination* (BILC). BILC is NATO's advisory and consultative body for language training and testing matters. BILC conducts a number of recurrent activities to ensure and enhance the quality of STANAG 6001 language assessments. For instance, BILC sponsors regular training and standardization seminars for STANAG 6001 testers and language teachers from NATO and Partner nations, organizes STANAG 6001 testing workshops and carries out consultative visits to assist nations in implementing a STANAG 6001-based language training and testing system. BILC promotes cooperation between member states for developing, piloting and moderating STANAG 6001 test items, thus furthering the common interpretation of the scale and the unified approach to language testing throughout the alliance. Last but not least, BILC instigated the development of the Benchmark Advisory Test, a NATO approved STANAG 6001 test of English that member nations can use for calibrating their own national STANAG 6001 tests.

In contrast, the CEFR does not have an international regulatory body, or any staff or official representatives. The Council of Europe, which commissioned the development of the CEFR framework, cannot play the role of a monitoring agency and it cannot hold users to account for or check out the validity of the claims about their tests. As an intergovernmental organization, its role is essentially political and it has neither the mission nor the resources to act in this way (Trim, 2012a: 19). Some limited guidance on good practice in implementing the CEFR was provided in the text itself (Council of Europe, 2001: 177-179) and in subsequent publications (e.g., Trim, 2001; Council of Europe, 2009; ALTE, 2011). However, since there is no mechanism for enforcing standardisation of judgements in relation to the CEFR, it seems likely that where systems rely on unregulated qualitative decisions, divergent local norms will emerge (*cf.* Green, 2012: 88). According to Jones (2009: 7), it is unreasonable to assume that decisions on alignment made across the diverse linguistic and cultural landscapes of Europe will lead to the emergence of a coherent interpretation of the CEFR framework. Consequently, the CEFR does not provide a strong basis for claims of equivalence.

Another important repercussion is that there are no officially endorsed CEFR training courses or benchmark CEFR tests (Saville, 2012: 60). CEFR language certificates can be issued by anyone or any organization, whether or not the assigned levels are based on a validated proficiency test. STANAG 6001 language certificates, on the other hand, can only be issued by authorized, nationally recognized defence language testing centres. These testing centres adhere to high-quality standards to ensure that STANAG 6001 language certificates are recognized throughout NATO as a valid proof of language competence.

Empirical findings

Because of the practical and methodological challenges in doing a comparison of actual proficiency using STANAG 6001 and CEFR assessments, to date only a small number of empirical studies have addressed the comparability of the STANAG 6001 and CEFR frameworks. One study comparing the results of a group of examinees that took both the STANAG 6001 Benchmark Advisory Test and a CEFR test of reading comprehension showed that examinees who were assigned a Level B1 on the CEFR test attained a STANAG 6001 proficiency level of anything between Level 1 and Level 2+. Conversely, examinees who had achieved a STANAG 6001 Level 2 attained a CEFR level ranging from A2 to C1. The results at other levels showed a similar picture (Swender *et al.*, 2012: 134). These outcomes are corroborated by similar linking studies that included tests of both reading and listening comprehension (Buck *et al.*, 2008; Gratton and Di Biase, 2013).

The findings from these studies seem to confirm that no clear-cut correspondences between the two scales or between two tests can be drawn. How well the results from STANAG 6001 proficiency tests agree with results from tests based on the CEFR framework will depend on how well the tests being compared are aligned in purpose, test type, construct definition, test tasks, the type of language performances elicited, and scoring procedures (*cf.* Kenyon, 2012: 29; Clifford, 2012: 54; Green, 2012: 83).

The empirical and other linking studies further indicate that it might be possible to assign STANAG 6001 ratings to CEFR tests with a fitting test design, and vice versa. However, this entails a highly specialized and time-consuming standard-setting process, and any correspondences established will only apply to the actual test under consideration. The results cannot be generalized to other tests due to the lack of a standardized CEFR testing methodology or benchmark model. Therefore, a different CEFR test may yield different STANAG 6001 ratings, and vice versa.

Conclusions

The CEFR and STANAG 6001 are two valuable, but in essence self-contained frameworks, each with its own contexts of use and assessment methodology. The substantial differences between the two frameworks with regard to their purpose, the delineation of the proficiency levels, and the testing system preclude an accurate linkage. In particular, the level-specific TCA requirements, the threshold nature of the rating ('floor' and 'ceiling' performance) and the notion of a non-compensatory core make the STANAG 6001 approach to language testing so different from CEFR testing that the frameworks can only be compared to some extent and with limited precision, but they cannot be aligned.
Given the nature of the CEFR assessment development, a single crosswalk from the CEFR scale to the STANAG 6001 scale is deemed unattainable. Instead, a specific crosswalk may have to be created for each CEFR and STANAG 6001 test and validated for different populations and uses. However, since any established correspondences will be applicable only to the specific test under consideration, the practical use of such a crosswalk will be severely limited.

In addition, due to the absence of standardized CEFR testing protocols or a monitoring body, the quality of CEFR tests tends to vary greatly. As it is not usually known on which interpretation of the CEFR level descriptors and on what type of assessment a CEFR rating is based, it is hard to establish the true significance of a CEFR language certificate. This inherent uncertainty is one of the reasons that NATO only accepts language certificates that are based on an official STANAG 6001 test and issued by a nationally recognized defence testing centre (*cf.* NATO SHAPE Directive 75-4; NATO ACO Directive 45-1).

In the current state of research, BILC holds the position that language certifications acquired on a CEFR test should not be transferred in terms of STANAG 6001 levels, or vice versa. Consequently, BILC does not recognize any conversion table or chart claiming correspondences between the two frameworks. Due to the incompatibility of the two frameworks, any comparison of levels and ratings will at best result in rough approximations. However, rough approximations of competencies are not helpful to either the employer or the individual. Overestimation of language abilities can easily lead to the assignment of duties that employees cannot cope with linguistically. Underestimation of language competence will give insufficient credit for the true abilities of the individual, which in turn may harm the individual's career opportunities.

Recommendation

BILC acknowledges the importance of using previously acquired language competencies in civilian contexts. But rather than trying to convert test scores and attained levels, which will ultimately be no more than just rough and unreliable approximations of the individual's true language competence, both retirees and future employers are better served by a proper, meaningful certification through a valid language test based on the appropriate language scale.

Nations that feel the need to formally recognize the foreign language proficiency of personnel leaving the defence organization are advised to make funds or other means available for recertification. In most instances, civilian educational institutions, government bodies and employers only accept language certificates from official, internationally recognized language exams, and not conversions of ratings based on a scale or test with which they are not familiar. Conversely, there is no requirement to convert civilian certifications into STANAG 6001 levels, because for international NATO assignments, only language certificates are accepted that are based on STANAG 6001 tests.

As a final point, the Working Group does not consider the issue of portability of STANAG 6001 certification into civilian contexts to be a concern of NATO or BILC. Whether and how a nation transfers STANAG 6001 ratings into civilian qualifications is essentially a national responsibility, although BILC strongly discourages conversion or recognition of test results acquired on a test based on one framework in terms of levels from another framework.

LIST OF WORKS CONSULTED

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.

Buck, G., Papageorgiou, S., and Platzek, F. (2008). *Exploring the Theoretical Basis for Developing Measurement Instruments on the CEFR.* Presentation at the EALTA Conference, Athens, 2008.

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. and Cox, T.L. (2013). Empirical Validation of Reading Proficiency Guidelines. *Foreign Language Annals*, 46, 1: 45-61.

Council of Europe (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment.* Retrieved 30 June 2017, from www.coe.int/lang-CEFR.

Council of Europe (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages.* Retrieved 30 June 2017, from www.coe.int.

Gratton, F. and Di Biase, M.J. (2013). Comparative Study between CEFR and STANAG – a case study conducted in two language institutes in Perugia, in: *Proceedings of the XVI AICLU conference at the University of Perugia (Italy)*, Perugia: Guerra Ed., 19-34.

Green, A. (2012). CEFR and ACTFL Crosswalk: A Text Based Approach, in: Tschirner, E. (ed.), *Aligning Frameworks of Reference in Language Testing.* Tübingen, Stauffenburg Verlag, 83-92.

Jones, N. (2009), A comparative approach to constructing a multilingual proficiency framework: Constraining the role of standard setting. *Cambridge ESOL Research Notes*, 37: 7-9.

Kenyon, D. (2012). Using Bachman's Assessment Use Argument as a Tool in Conceptualizing the Issues Surrounding Linking ACTFL and CEFR, in: Tschirner, E. (ed.), *Aligning Frameworks of Reference in Language Testing.* Tübingen, Stauffenburg Verlag, 23-34.

Little, D. (2012). Elements of L2 Proficiency: The CEFR's Action-Oriented Approach and Some of its Implications, in: Tschirner, E. (ed.), *Aligning Frameworks of Reference in Language Testing.* Tübingen, Stauffenburg Verlag, 71-82.

Lowe, P. Jr. (2012). Understanding "Hidden Features" of the ACTFL Speaking Guidelines as an Intermediate Step to Comparing the ACTFL Guidelines and the CEFR for Speaking Assessment, in: Tschirner, E. (ed.), *Aligning Frameworks of Reference in Language Testing.* Tübingen, Stauffenburg Verlag, 93-106.

NATO (North Atlantic Treaty Organization) – Military Agency for Standardization (1976). *NATO Standardization Agreement (STANAG) 6001: Language Proficiency Levels. Edition 1.* Retrieved 12 June 2017, from www.natobilc.org.

NATO (North Atlantic Treaty Organization) - Standardization Office. (2016). *ATrainP-5. NATO STANAG 6001: Language Proficiency Levels. Edition A Version 2.* Retrieved 12 June 2017, from www.natobilc.org.

NATO (North Atlantic Treaty Organization) – Allied Command Operations. *ACO Directive No. 45-1 – ACO Military Personnel Management and Administration for PE Posts, Chapter 3 – Language Skills*. Issued 13 April 2015.

NATO (North Atlantic Treaty Organization) – Supreme Headquarters Allied Powers Europe. *SHAPE Directive No. 75-4 – Language Testing*. Issued 9 May 2006.

North, B. (1993). *The Development of descriptors on scales of proficiency: perspectives, problems, and a possible methodology.* NFLC Occasional Paper, National Foreign Language Center, Washington D.C.

North, B. (2014). *The CEFR in Practice.* Cambridge, UK: Cambridge University Press.

Saville, N. (2012). The CEFR: An Evolving Framework of Reference, in: Tschirner, E. (ed.), *Aligning Frameworks of Reference in Language Testing.* Tübingen, Stauffenburg Verlag, 57-70.

Saville, N. and Gutierrez Eugenio, E. (2016). The European Commission's 'Study on comparability of language testing in Europe' (2015). *Cambridge ELA Research Notes*, 63: 3-12.

Swender, E., Tschirner, E. & Bärenfänger, O. (2012). Comparing ACTFL/ILR and CEFR Based Reading Tests, in: Tschirner, E. (ed.), *Aligning Frameworks of Reference in Language Testing.* Tübingen, Stauffenburg Verlag, 123-138.

Trim, J. (ed.). (2001). *CEFR. Guide for Users*. Strasbourg, Council of Europe. Retrieved 30 June 2017, from www.coe.int.

Trim, J. (2012a). Provo Address, in: Tschirner, E. (ed.), *Aligning Frameworks of Reference in Language Testing.* Tübingen, Stauffenburg Verlag, 19-22.

Trim, J. (2012b). Some earlier developments in the description of levels of language proficiency, in: Green, A. (ed.), *Language functions revisited. Theoretical and empirical bases for language construct definition across the ability range*. Cambridge, UK: Cambridge University Press, xxi-xli.