Rethinking the Rating Process: Solution to the Threshold Performance Dilemma

Mary Jo DiBiase-Lubrano, Yale University (USA)
Jana Vasilj-Begovic, Department of National Defence (CAN)

Abstract

Test reliability and intrinsically, validity are indisputably linked to valid scoring criteria. When assessing constructed responses in very advanced language proficiency level tests specifically, the extent of the raters' level of training and norming to the criteria, and their ability to foresee all possibly correct responses are critical. These issues have a bearing in standard setting methods aiming to establish two meaningful categories of candidates who meet and do not meet the minimum criterial levels of performance, which exemplify the construct and reflect the test purpose. This conceptual paper describes an approach to establishing cut scores in an integrated reading comprehension test via the skill of writing, which purports to measure Distinguished proficiency reading skills via constructed responses. A mixed method approach is adopted in which raters' holistic ratings are triangulated with their analytic scores. The method, which the authors called the Retrodictive Modeling Approach (RMA), relies on raters' level of expertise and holistic ratings coupled with their qualitative analysis and scores, which yield a pattern useful for establishing a threshold performance level. Although claims to generalizability are beyond the scope of this study, further research may lead to a wider use of the RMA in Distinguished proficiency level testing.

Keywords: standard setting; norming; scoring validity, Distinguished levels

Very advanced proficiency testing, such as the Distinguished level on the American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines or Level 4 on the Interagency Language Roundtable (ILR) proficiency scale presents unique challenges given the limitations of the test design, the selection of appropriate item tasks, the level of rater training, and the choice of appropriate scoring techniques. Nonetheless, while the usefulness of testing Distinguished or Level 4 proficiency levels is beyond the scope of this study, there are contexts and job descriptions which entail the use of highly articulated and nuanced language. One such context is language proficiency in specific headquarters within the North Atlantic Treaty Alliance (NATO), where military and civilian job descriptions call for STANAG Level 4 proficiency in one or more skills (LNA, 2015). To address the complexities of testing at such high levels, the Bureau of International Language Coordination (BILC), NATO's advisory body for language training and testing matters, was tasked with establishing a working group to develop a prototype of a reading comprehension test at Level 4 according to the proficiency scale in use among NATO countries, the Standardization Agreement No. 6001. Since testing at STANAG 6001 Level 4 was groundbreaking, the working group agreed to start from the skill of reading, because it was considered to be the easiest of the four skills to measure. The purpose of this test was to measure the reading proficiency at STANAG 6001 Level 4, while the use of the test scores would be to certify the reading proficiency of military and civilians whose positions entailed such a high level of literacy. Although testing is considered a national responsibility, BILC assists nations with test development by promoting best practices and by fostering a uniform interpretation of the STANAG 6001 language proficiency descriptors, used as a guideline for curriculum and test development, as well as for the construct of general proficiency. The ultimate goal of BILC's efforts is to contribute to interoperability within NATO's multi-national operations and deployments, contexts in which English is used as the main language of communication, and at the same time, recognized as a significant impediment to interoperability. The STANAG 6001 has been in use since 1976 (Edition 1), and constitutes the underpinning for standardized testing practices and procedures in NATO, Partnership for Peace countries (PfP), and other global partner nations that wish to certify their personnel with linguistic profiles based on this standard. While the document has been updated several times since 1976, the changes only regard the preface and instructions to participating nations, whereas the actual proficiency descriptors have remained unchanged since its 2003 edition.

To carry out this project, BILC formed a working group (WG), selected on members' level of expertise in working with the scale, and their documented background as test developers in their country of residence (Bulgaria, Canada, Denmark, Germany, the Netherlands, Sweden, and the USA). The group was tasked

with the development of an advisory/prototype test starting with the skill of reading, defined in STANAG 6001 as the expert level and described as:

"Demonstrates strong competence in reading all styles and forms of the written language used for professional purposes, including texts from unfamiliar general and professional-specialist areas. Contexts include newspapers, magazines, and professional literature written for the well-educated reader and may contain topics from such areas as economics, culture, science, and technology, as well as from the reader's own field. Can readily follow unpredictable turns of thought on any subject matter addressed to the general reader. Shows both global and detailed understanding of texts including highly abstract concepts. Can understand almost all cultural references and can relate a specific text to other written materials within the culture. Demonstrates a firm grasp of stylistic nuances, irony, and humour. Reading speed is similar to that of a native reader. Can read reasonably legible handwriting without difficulty." (STANAG 6001, 2016, A-6).

STANAG 6001 is a criterion-referenced scale in which each descriptor can be considered a separate construct (Clifford, 2013 personal communication) for test development purposes in a particular skill, and for a particular proficiency level. All STANAG 6001 descriptors contain information that relates to the content domains, language tasks, accuracy statements and text types, i.e. length and organization of text. The latter two inform the fundamental rating criteria used in the productive skills, whereas in the receptive skills, the statements serve to describe the nature of the written or audio passages the language user is able to comprehend. Thus, a proficiency level can represent a separate construct which contains all necessary components for test construction such, as domains, language tasks, accuracy, and text length. For Level 4 reading, the breakdown of the descriptor is illustrated below where content refers to the topical domains, as well as sources of higher-level texts; the linguistic tasks are the operations the reader needs to perform in order to process the text, while accuracy defines how well (conditions) the reader can understand the text. The text type, not specifically mentioned in the trisection, is understood to be of essay length discourse (STANAG 6001, 2016, A-6).

CONTENT

All styles and forms of writing used for professional purposes, including texts from unfamiliar general and professional-specialist areas.

Newspapers, magazines, and professional literature written for well-educated native readers.

Highly abstract concepts.

Reasonably legible handwriting.

TASKS

Follow unpredictable turns of thought on any subject matter addressed to the general reader.
Show both global and detailed understanding of texts.
Understand almost all cultural references.

ACCURACY

Can relate a specific text to other written materials in the culture.
Demonstrates a firm grasp of stylistic nuances, irony, and humor.
Shows both global and detailed understanding of texts.
Understands almost all cultural references.
Reading speed is similar to that of a native speaker.

Literature Review

Clifford (2013, personal communication) defines reading comprehension as "the active, automatic, far-transfer process of using one's internalized language and culture expectancy system to efficiently comprehend an authentic text for the purpose for which it was written." He states that the reader at Level 4 understands meaning "beyond the lines." In other words, the reader is able to make evaluative judgements or express opinions about a text; evaluate the significance of the author's message, credibility, intent, and purpose; extrapolate beyond the text; and place it in a socio-cultural and historical context. Evaluative comprehension involves making a judgement about the text genre and mode of discourse, rhetorical organization of the text, and the writer's use of jargon, figures of speech, and allusions. Evaluative comprehension skills entail taking into account unstated assumptions in order to understand, evaluate, accept or reject the writer's arguments. While understanding "between the lines" is a required skill for both STANAG 6001 levels 3 and 4, understanding "beyond the lines" is a skill that distinguishes Level 4 from lower proficiency levels (STANAG 6001). It is also important to mention that attaining Distinguished or higher levels of language proficiency implies not only possessing outstanding language skills, but also higher order thinking skills, such as deductive and inductive reasoning, analysing, and synthesizing. Level 4 readers understand highly sophisticated written target language on unfamiliar general, abstract or professional-specialist topics. They generally understand specialized language on general abstract and complex topics outside of their area of expertise (e.g., philosophical essays written for a well-educated general reader). These readers can discern relationships among sophisticated written materials in the context of broad experience, and can follow unpredictable turns of thought in editorial, conjectural, and literary materials in any subject matter intended for the general reader. They

also recognize, understand and usually correctly interpret cultural allusions, nuance and emotional overtones and attitudes (disenchantment, satire, humour, etc.) as well as less common figures of speech, such as malapropisms and spoonerisms.

Child (1998) elaborated on text difficulty and added that an author's unique point of view, and the method of argumentation may be complex and innovative at higher levels of proficiency. According to Lowe (1998), Level 4 texts are "abstract & culturally dense, often have embedding syntax used with virtuosity," while Edwards (1996, 19) states that the author "may take a novel or creative approach to a problem" adding that "in these texts, the reader is likely to encounter highly individualized or culture-specific forms of discourse, abstract metaphors, and symbolism. The author assumes a great deal of reader input and leaves historical, cultural or other references and assumptions unexplained.

Full comprehension of such texts goes beyond the literal comprehension of explicitly stated information. It involves interpretation and critical evaluation of the text and the authorial intention. By evaluating the text, the readers "enter into a dialogue with the text and make the text their own" (Alschuler et al, 2002, 6).

As valid and reliable testing involves accurate elicitation and rating procedures (Alschuler et al, 2002), the working group felt that only constructed response (CR) task types would be appropriate to elicit evidence of reading comprehension. At this level, multiple-choice questions, as a testing technique might become a puzzle or a riddle-solving exercise as the options would tend to become exceedingly convoluted, and too close in meaning to each other. Justification for adopting CR was also found in Shrock & Coscarelli's (2007) analysis of the cognitive abilities associated with proficiency levels (Figure 1) and the corresponding appropriateness of task types.

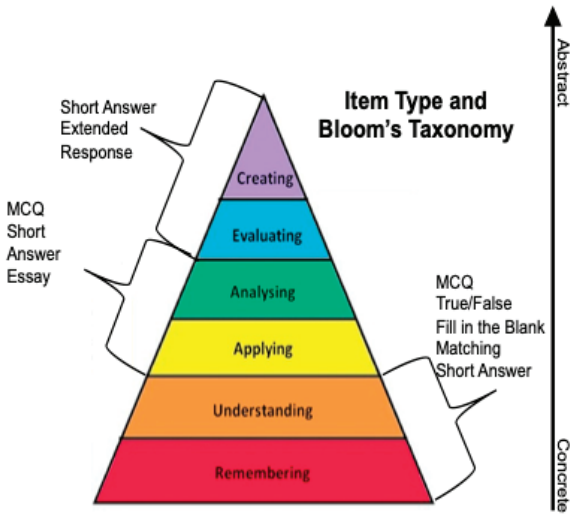

Fig. 1: Task types and cognitive abilities (Shrock & Coscarelli, 2007)

However, using CR task types at this level poses challenges in determining a suite of acceptable responses at a level where "reading entails a cognitive process which involves not only language and reading skills, but also general intellectual reasoning" (Aschuler et al, 2002). Additionally, in these reading tasks, "shared information and assumptions are at a minimum (and) personal input is paramount" (Child, 1998, 6). This places exponentially increased demands on raters (Child, 1998, 22).

Another layer of challenge lies in the possibility of test-takers' offering a variety of new interpretations of the text that may be plausible and thereby acceptable. In turn, raters themselves tend to have various interpretations of the test-takers' responses if the responses fall out of the fully acceptable or fully unacceptable ranges.

To address the complexities of rating CR at Level 4 and to determine a minimally acceptable performance or threshold level on the reading-comprehension prototype test, the WG adopted a novel standard setting approach the authors called the Retrodictive Modelling Approach (RMA), in which raters' conceptual understanding of Level 4 proficiency was triangulated with the patterns which emerged from their analytical ratings. These sessions were conducted in plenary and online using email and Qualtrics surveys. The authors, as group leaders, were responsible for collating and analyzing the patterns. In turn, the cut score yielded was further validated through multiple rounds of ratings where samples were recoded and anonymized, and qualitative evidence in the form of surveys collected from the raters.

Most standard setting methods rely on the expertise of the "judges" involved, in terms of familiarity with the test, understanding of the inferences and decisions made on the test scores, and conceptualization of how minimal criteria of performance might be exemplified (Angoff, 1971).

Kane (1998), claims that there is no real cut score to be found, but that rather the cut score needs to be created on the basis of many assumptions, the most important among these being the test purpose. Notwithstanding, determining the cut score may be fraught with issues stemming from the validity and reliability of the test itself (Dwoning et al, 1997), the examiners, (Schmitt et al, 1990), e.g. leniency/strictness, etc., the examinees, e.g. internal and external reliability issues deriving from concentration, health, psychological state, etc. Very similarly, the principles underpinning the RMA method lie on the assumption that the 'judges' are well-versed with the STANAG 6001 proficiency descriptors, understand the inferences made on the test scores, and are able to conceptualize what a threshold performance would look like in real world scenarios.

Statement of the Problem and Research Questions

As testimony to the challenges of establishing meaningful categories of those who meet the standards and those who do not, is the fact there is no "golden" standard setting method but rather a suite of over 400 methods that stakehold-

ers may choose from for the purposes of their test project. Most methods focus on establishing pass/fail judgements (Cizek et al, 2007). Similarly, the RMA aims to determine two categories of test takers using empirical evidence as to whether individual responses to items meet the criteria at Level 4 with respect to judges' evaluations of successful, unsuccessful or partial responses in combination with qualitative judgments, based on a holistic evaluation of the test taker's performance being at Level 4 or not. The analysis of both approaches will yield a comparison of a patterned profile made up of the minimal number of successful (S), unsuccessful (U) and partial (P) responses which consistently correspond to the judges' holistic evaluation of a passing performance.

The RMA approach was applied to the STANAG 6001 Level 4 reading - to - write prototype test and its effectiveness was trialled on sample papers submitted by a selected cohort of test takers, thought to be Level 4 readers based on evaluative judgments made by teachers, currently held positions requiring such level, or prior national test scores. This paper intends to answer the following questions about the RMA approach, and its effectiveness to establish cut scores on a CR test in order to assess Distinguished levels of reading proficiency, in accordance with STANAG 6001 Level 4.

RQ1: *Does the threshold performance, identified through the RMA, accurately reflect the construct of STANAG 6001 Level 4 reading prototype test and validity of uses and interpretations that can be made on the basis of the test scores?*

RQ2: *How do judges' holistic evaluations correlate with their analytical scores, and how reliable are they in predicting the threshold performances*?

Method

Test Design and Development

The testing format selected was an integrated skills approach, i.e., reading tested via writing. Considering the challenges of choosing the appropriate test item type discussed earlier, a CR would enable the test taker to provide answers to complex questions, which characterize Level 4 reading texts, and which require the use of higher order thinking skills. Although the skill of writing was not tested *per se*, the WG felt that test takers would need to possess level 3 writing proficiency at a minimum, to be able to demonstrate comprehension of Level 4 texts.

The test format is thus a single-level test comprised of two authentic reading texts, each approximately 1400 words long. One text deals with a topic from the military doctrine field, and the other with a social/political issue. Both texts are intended for general educated readers, i.e., with no particular special area of expertise, or educational background. Each text is followed by six items, written in English and testing different Level 4 reading tasks, as per the STANAG 6001 descriptor,

e.g., understand the author's choice of words in relation to the tone and persuasive stance/attitude, link ideas from various parts of the text, interpret the text in its wider cultural and societal context, follow unpredictable turns of thought, understand cultural references and allusions, etc. Paragraphs in each text are numbered, as tasks/questions may refer to a particular paragraph or to the text as a whole. One example of a Level 4 text and task is given below:

Excerpt from "Politics among Nations" *by Hans J. Morgenthau*

[Para 1] When one reflects upon the development of American thinking on foreign policy, one is struck by the persistence of mistaken attitudes that have survived- under whatever guises - both intellectual argument and political experience.

[Para 2] Once that wonder, in true Aristotelian fashion, has been transformed into the quest for rational understanding, the quest yields a conclusion both comforting and disturbing: we are here in the presence of intellectual defects shared by all of us in different ways and degrees. Together they provide the outline of a kind of pathology of international politics. When the human mind approaches reality with the purpose of taking action, of which the political encounter is one of the outstanding instances, it is often led astray by any of four common mental phenomena: residues of formerly adequate modes of thought and action now rendered obsolete by a new social reality; demonological interpretations of reality which substitute a fictitious reality - peopled by evil persons rather than seemingly intractable issues - for the actual one; refusal to come to terms with a threatening state of affairs by denying it through illusory verbalization; reliance upon the infinite malleability of a seemingly obstreperous reality.

[Para 3] Man responds to social situations with repetitive patterns. The same situation, recognized in its identity with previous situations, evokes the same response. The mind, as it were, holds in readiness a number of patterns appropriate for different situations; it then requires only the identification of a particular case to apply to it the preformed pattern appropriate to it. Thus, the human mind follows the principle of economy of effort, obviating an examination de novo of each individual situation and the pattern of thought and action appropriate to it. Yet when matters are subject to dynamic change, traditional patterns are no longer appropriate; they must be replaced by newness reflecting such change. Otherwise, a gap will open between traditional patterns and new realities, and thought and action will be misguided.

Sample Test Question (Whole Text): In the realm of international politics, what intellectual failings and attitudes do the Americans exhibit, and what are their origins?

Sample Response: When it comes to foreign policy and international politics the Americans tend to rely on obsolete views and allow their action to be guided by the principles that are no longer applicable to the newly arisen circumstances. Instead of adapting their views to the new realities and examining every new political situation with a fresh view, they misinterpret it by viewing it through the prism of either illusory perception or substituting that reality.

No specific instructions are given to the test takers as to the expected length of the response, except that it should be as complete, concise and coherent as possible. During the rating procedure, it was observed that some responses, albeit cryptic still showed clear evidence of comprehension, while some longer ones showed the opposite. Access to a pen and paper for note-taking was part of the administration procedure.

A complete guide to rating, including sample responses, was provided to all raters involved in the RMA piloting.

The Retrodictive Modelling Approach (RMA)

The WG members were experienced test developers and experts, who were all well-versed with the STANAG 6001 scale proficiency descriptors. Their professional background within the military context ensured full understanding of the inferences which would be made on test scores used for the purposes of this advanced proficiency level test. Two members of the working group had also participated in the development of this test, whereas the others had moderated or revised the multiple versions of the test before official piloting.

A tester booklet, containing the texts, administrative notes and instructions for the test takers, the rating process, as well as the sample responses (answer key), was provided to all raters within the WG. Each sample response, as per the above given example, specified the Level 4 reading task it was testing and the paragraph it referred to. The first rounds of ratings of the piloted tests included awarding an initial holistic score based on the STANAG 6001 Level 4 descriptor, and a detailed analytical score, reported as "Successful", "Unsuccessful" and "Partially successful." The latter option was included for those responses which did not fully match the ones in the answer key, but that, nonetheless, demonstrated that the test taker comprehended the question, and therefore the text to some extent. Proficiency levels in criterion-referenced tests based on STANAG 6001 are ranges of language performances. Test takers need to demonstrate mastery of language tasks within the content domains, and the accuracy demands of the level. Compensatory lan-

guage behaviour, such as, for example, a test taker's aggregated and cumulative correct responses on a multi-level test cannot correspond to a proficiency level. Only a minimum number of correct responses at each level can determine mastery, or sustainment of performance.

To the extent that test developers strive to minimize the impact of extraneous, unpredictable and construct-irrelevant variables such as random reliability issues on test scores, these are nevertheless present (Davis, 2003).

Given that this reading test prototype is a mono-level test, and the broad spectrum of content domains and language functions it purports to measure in accordance with STANAG 6001 descriptors for reading comprehension, it is virtually impossible for test takers to get 100% correct answers on any given proficiency test (perhaps possible on an achievement test based solely on predictable curriculum content), raters decided to allow credit for partial answers to be factored in.

The scoring criteria, as outlined below, were developed as an analytical approach to rating. For the rating of the speaking and the writing skills, analytical ratings have been developed within the BILC community; however, none existed for the receptive skills. The intent with the criteria below was to design a simple analytical framework, pilot it and attempt to correlate the scores with the Level 4 descriptor by assigning a final holistic rating.

> **SUCCESSFUL** The examinee addressed the task and sub-tasks (if applicable) and he/her response either reflected fully the ideas expressed in the Sample response or provided another acceptable interpretation of the text. The response was sufficiently elaborate, precise and coherent.
>
> **PARTIAL** The examinee's answer did not fully address the task or it captured part of the basic idea(s) or details conveyed in the Sample Response. Even though the response was written with precise and coherent language, its content only partially corresponded to that given in the Sample Response.
>
> **UNSUCCESSFUL** The examinee's answer deviated from the information contained in the Sample Response showing a lack of comprehension of the text. It was written in a simple and concrete manner that did not convey the basic ideas contained in the Sample response.

Determining a minimally acceptable performance, or cut score was exceptionally challenging in this type of test, which integrated the two linguistic skills – reading and writing. Establishing two meaningful categories of those candidates who had met the criteria of sustained ability to understand a Level 4 reading passage via the modality of writing – even though their writing ability was not assessed per se'

– and those who did not was a fundamental aspect of the WG's tasks. The authors believe the RMA can be adopted to help establish these categories. The approach relies heavily on the expertise of the raters, who must be familiar with the STANAG scale, be familiar with the test itself, and with the inferences and decisions made on the test takers in conjunction with their inferred proficiency in the real world.

Given the high proficiency level being assessed, and the unique characteristics of Level 4 individualistic texts (Lowe,1998, 358), it is virtually impossible to anticipate all the possible responses candidates might be able to produce, which could nevertheless be evaluated as either partially or fully correct.

Flipping the rating process from the predictive to the retrodictive method offered raters a new assessment perspective and procedure. In order to establish the cut score, the working group believed that the five raters were the most suitable "judges" for the standard setting sessions (Angoff, 1971). Starting from a conceptualization of what the raters deemed a Level 4 reader could do, albeit minimally and correlating this evaluation with a more analytical assessment of the individual items was thought to yield a minimum profile of a threshold performance, which included the least number of successful, partial and unsuccessful responses, and which reflected the construct of Level 4 reading comprehension. The raters created this approach, the Retrodictive Modelling Approach (RMA), which aims to exemplify the criterial level of performance reflecting the Level 4 construct; it is a mixed, evidence-based rating process through which performances are assessed in three rating steps. The three rating steps needed to establish rating patterns, which would yield a minimum cut score, are outlined below:

- Step 1 - **Initial Holistic Rating (IHR)**: Raters read all responses globally, compare their overall impression of performance against the STANAG 6001 Level 4 descriptor, and mark their rating as "At Level" or "Below Level" for each individual item/task. The assumption is that all raters involved in this process have a clear theoretical concept of a Level 4 reader – at least holistically;
- Step 2 - **Analytic Rating-(AR)**: Raters analyse the responses individually by comparing them against the sample responses (answer key), and marking them as **Successful (S), Partial (P),** or **Unsuccessful (U)**;
- Step 3 - **Final Holistic Rating (FHR)**: Raters decide whether the test is a **PASS** at Level 4 by awarding it a final holistic rating that reflects the ranges of reading proficiency in accordance with STANAG 6001 Level 4 descriptor.

All tests were coded and the samples rated by five judges. When the initial holistic rating matched the final holistic rating, for example, a rating of "pass" (i.e.

candidate is deemed a Level 4 proficient reader), the authors analysed the analytical ratings to determine patterns of ratings, which emerged in terms of a specific number of successful, unsuccessful, and partial evaluations. The minimum number of "successful" scores for this profile, which consistently represented a match between the IHR and FHR would tentatively reflect the cut score.

The objective of the RMA-based rating method is to assist raters in the qualitative conceptualization of Level 4 reading proficiency in accordance with STANAG 6001. Employing this approach and forming a conceptual model of Level 4 (Distinguished) proficiency also enabled the WG to develop *a posteriori* rating criteria and establish the MAC.

The RMA procedure included:

1. norming sessions to agree on text and item levels (calibration);
2. norming sessions to agree on acceptable responses;
3. matching responses to the descriptor;
4. identifying patterns of performances at different ranges within the same level (threshold, mid, high);
5. analysing raters' scores on the performance grids;
6. trialling the scoring on actual responses.

The model retrospectively analyses patterns in the analytical scores, which correspond to raters' holistic evaluation of Level 4 (Distinguished) proficiency. The minimum common denominators, which are a combination of the Successful (S), Partial (P), and Unsuccessful (U) ratings or scores ultimately yield a minimally acceptable profile, which still corresponds to the WG's conceptualization of a threshold Level 4 reader.

Participants

The test was piloted in three phases, on a total of 38 candidates, a mix of military (N=32) and civilians (N=6), native (N=3) and non-native readers (N=35). The main piloting which involved 18 test takers took place at the International Military Staff (IMS), NATO Headquarters in Brussels, Belgium. All trialled test papers were rated remotely in three rounds: (a) an initial round during which a preliminary calibration and validation of the answer key were established; (b) a second, where the raters rated independently, and then submitted their results to the authors, and (c) a third, final round, where the established cut score was applied in order to validate its effectiveness in reflecting the construct. All samples were anonymized and

labelled "sample A," "sample B," etc for each round of ratings, and they were also typed to facilitate reading.

A survey was created so that raters could record their evaluations. A box for comments was made available at the end of each sample for raters to add comments and observations about the samples, the ratings or the procedure, including the initial holistic score and its confirmation after having rated analytically.

Trialling the test included recruiting volunteers who were on active duty in international positions, which had been labelled by NATO stakeholders as requiring Level 4 proficiency. Local proctors ensured that the test was trialled in secure environments and specific instructions were given to both the test takers and the proctors. A test control officer from Belgium administered the test at NATO HQ to 15 test takers, and the Chief of Testing at SHAPE administered the test to 5 test takers. Subsequent trialling was conducted on 18 test takers, out of whom three were located at Supreme Allied Command Transformation (SACT), Norfolk, four at the Defence Military College in Sweden, three at the Defence Military College in Denmark, three at the Canadian Defence Academy, one at the Netherlands Defence Academy, three at Bundessprachenamt, Germany, and one at a private university on the East Coast, USA.

The rating process of the initial trials included applying the RMA process. In the meantime, test items/tasks were modified based on the analysis of the initial 18 test administrations. Raters were asked to provide an Initial Holistic Rating (IHR), and then proceed to rating the individual items (N=12) of the two texts. Subsequently, a Final Holistic Rating (FHR) would either confirm the initial holistic rating, or reject it given the analytic evidence of the individual items evaluated.

Results were compiled and a tentative cut score of 7 Successful 5 Partial and 2 Unsuccessful responses emerged that consistently reflected raters' understanding of STANAG Level 4 reading proficiency at its threshold/minimum level. Once the tentative cut score had been identified, an additional round of ratings was conducted by applying the newly established cut score (7S, 5P and 2U). Raters followed the above-described three-step rating process: 1) provide an IHR; 2) rate analytically; 3) provide a FHR. A box for comments was provided in which raters could elaborate on whether their holistic rating had changed based on the cut score.

Statistical tests and analysis

A FACETS analysis was run by considering the three facets of examinees, raters and test items. In order to do this, all Successful, Unsuccessful and Partial responses were converted to 2, 0 and 1, respectively. Similarly, the twelve examinee samples were coded A, B, C, D, G, J, K, L, N, R, S, T while the names of the raters were referred to as, rater 1, rater 2, rater 3, rater 4 and rater 5. Although the actual test comprised of two texts with six items each, the authors considered twelve items in the analysis as the initial holistic rating and the final holistic rating were

each considered as additional items, for a total of 14 items. They were coded a T1 (referring to first of the two texts) and 1 (referring to the first item of that text so that T1_1 would refer to the first item of text 1 and T2_2 would refer to the second item of the second text and so on).

The examinees were considered as the "floating" facet whereby the greater the score, the greater the ability of the examinee. Instead, the rater facet analysis was run to determine the inter and intra rater reliability. An initial observation of the *fit* statistics shows the variability in rater harshness or leniency with a strong indication of rater self-consistency. This would lead to the importance of always having two raters in the rating process, which is consistent with best practices in test development projects.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Exact Agree. PtExp	Obs %	Exp %	N Raters
147	144	1.02	1.07	-.53	.12	1.22	1.9	1.37	1.9	.63	.50	.63	59.0	53.7	5 rater 5
134	144	.93	.93	-.35	.12	.88	-1.1	.80	-1.1	1.11	.68	.63	68.6	55.3	2 rater 2
112	144	.78	.69	-.03	.12	.92	-.6	.81	-1.0	1.19	.68	.64	70.1	57.0	3 rater 5
105	144	.73	.62	.08	.12	.82	-1.5	.67	-1.8	1.28	.72	.63	76.2	57.2	4 rater 4
61	144	.42	.25	.84	.14	1.26	1.7	1.17	.6	.74	.47	.58	62.8	55.5	1 rater 1
111.8	144	.78	.71	.00	.13	1.02	.1	.96	-.3	.61			Mean (Count: 5)		
29.5	.0	.21	.28	.47	.01	.18	1.5	.26	1.4	.10			S.D. (Population)		
33.0	.0	.23	.32	.53	.01	.21	1.7	.29	1.5	.12			S.D. (Sample)		

Model, Populn: RMSE .13 Adj (True) S.D. .46 Separation 3.65 Strata 5.20 Reliability (not inter-rater) .93

Model, Sample: RMSE .13 Adj (True) S.D. .52 Separation 4.11 Strata 5.81 Reliability (not inter-rater) .94

Model, Fixed (all same) chi-square: 63.3 d.f.: 4 significance (probability): .00

Model, Random (normal) chi-square: 3.8 d.f.: 3 significance (probability): .29

Inter-Rater agreement opportunities: 1440 Exact agreements: 970 = 67.4% Expected: 802.5 = 55.7%

The *fit* statistics of the test items are also within parameters although the items are reliably different from each other. There is evidence that the items might range in terms of "difficulty" or "ease". This finding is consistent with the conceptual framework of the STANAG 6001 scale whereby each level is a threshold level within a range of proficiency which is wider as the scale progresses, e.g. level 1 range is much narrower compared to the Level 4 range in terms of content, task, accuracy expectations and text type. The easiest item is the third item of text 2, with a logit

of -1.52 whereas the harshest is item 2 of text 2 with a logit of 1.05. Furthermore, the final holistic rating was more stringent than the initial holistic rating, indicating that the analytical scores led to a negative final holistic score.

Total Total Obsvd Fair(M)| - Model | Infit Outfit |Estim.| Correlation | |
Score Count Average Average|Measure S.E. | MnSq ZStd MnSq ZStd |Discrm| PtMea PtExp | Nu Text_Item |
88 60 1.47 1.63 | -1.40 .19 | 1.18 1.0 1.25 .7 | .86 | .41 .51 | 9 T2_3
73 60 1.22 1.34 | -.89 .18 | 1.12 .8 1.00 .1 | 1.28 | .57 .57 | 5 T1_5
56 60 .93 .93 | -.36 .18 | .83 -1.0 .79 -.8 | 1.06 | .67 .61 | 7 T2_1
56 60 .93 .93 | -.36 .18 | .81 -1.2 .65 -1.5 | 1.53 | .74 .61 | 11 T2_5
48 60 .80 .74 | -.10 .18 | .71 -1.8 .79 -.7 | .68 | .64 .61 | 12 T2_6
45 60 .75 .66 | .01 .19 | 1.45 2.2 1.60 2.0 | .43 | .36 .61 | 6 T1_6
44 60 .73 .64 | .04 .19 | .73 -1.5 .69 -1.2 | 1.17 | .71 .61 | 1 T1_1
42 60 .70 .59 | .11 .19 | 1.18 1.0 1.37 1.2 | .80 | .51 .61 | 3 T1_3
33 60 .55 .40 | .45 .20 | 1.13 .6 1.23 .7 | .84 | .50 .59 | 4 T1_4
29 60 .48 .33 | .62 .21 | 1.14 .7 .90 -.1 | 1.00 | .55 .58 | 2 T1_2
26 60 .43 .27 | .76 .22 | 1.10 .5 .89 -.1 | .93 | .52 .56 | 10 T2_4
19 60 .32 .18 | 1.12 .24 | .75 -1.0 .39 -1.4 | 1.41 | .69 .51 | 8 T2_2
46.6 60.0 .78 .72 | .00 .19 | 1.01 .0 .96 -.1 | | .57 | Mean (Count: 12)
18.9 .0 .32 .42 | .67 .02 | .22 1.3 .33 1.1 | | .11 | S.D. (Population)
19.8 .0 .33 .43 | .70 .02 | .23 1.3 .34 1.1 | | .12 | S.D. (Sample)

Model, Populn: RMSE .20 Adj (True) S.D. .65 Separation 3.30 Strata 4.74 Reliability .92
Model, Sample: RMSE .20 Adj (True) S.D. .68 Separation 3.46 Strata 4.95 Reliability .92
Model, Fixed (all same) chi-square: 133.6 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.2 d.f.: 10 significance (probability): .43

The *fit* statistics of the examinees shows that most are within a normal range of .5 – 1.5. Examinee A is an outlier as their *fit* statistics are 1.83: at a closer look, raters do not seem to follow the pattern in using the 0 rating (i.e. unsuccessful). Overall, the reliability is .96 which is a strong indicator that the examinees are at different levels of ability. Since this is a mono-level test, it might point to the difficulty of the actual items.

Total Total Obsvd Fair(M)| + Model | Infit Outfit |Estim.| Correlation |
| Score Count Average Average|Measure S.E. | MnSq ZStd MnSq ZStd |Discrm| PtMea PtExp | Nu Examinees |
| 16 60 .27 .17 | -1.58 .24 | 1.02 .1 .68 -.5 | 1.30 | .54 .41 | 7 K
| 21 60 .35 .24 | -1.32 .22 | .84 -.6 .84 -.2 | .72 | .37 .45 | 10 R
| 25 60 .42 .30 | -1.14 .21 | .56 -2.4 .52 -1.4 | .96 | .62 .47 | 12 T
| 26 60 .43 .31 | -1.09 .20 | .86 -.6 .70 -.8 | 1.29 | .60 .47 | 11 S
| 30 60 .50 .38 | -.94 .19 | 1.29 1.4 1.33 1.0 | 1.22 | .46 .49 | 4 D
| 35 60 .58 .47 | -.76 .19 | .86 -.7 .86 -.4 | .91 | .53 .51 | 8 L
| 38 60 .63 .53 | -.66 .18 | .90 -.6 .95 .0 | .72 | .47 .52 | 2 B
| 44 60 .73 .65 | -.47 .18 |1.31 1.9 1.31 1.2 | 1.26 | .47 .53 | 9 N
| 66 60 1.10 1.14 | .18 .17 | .95 -.2 .82 -.8 | 1.74 | .67 .54 | 5 G
| 73 60 1.22 1.29 | .38 .17 | .82 -1.2 .86 -.5 | .44 | .48 .53 | 6 J
| 92 60 1.53 1.65 | 1.02 .20 1.89 3.8 1.89 2.2 | .79 | .15 .47 | 1 A
| 93 60 1.55 1.67 | 1.06 .20 |.77 -1.2 .80 -.5 | .63 | .42 .47 | 3 C
|
| 46.6 60.0 .78 .73 | -.44 .20 | 1.01 -.1 .96 -.1 | | .48 | Mean (Count: 12)
| 26.2 .0 .44 .53 | .86 .02 | .33 1.6 .36 1.0 | | .13 | S.D. (Population)
| 27.4 .0 .46 .55 | .89 .02 | .35 1.7 .37 1.1 | | .14 | S.D. (Sample)
Model, Populn: RMSE .20 Adj (True) S.D. .83 Separation 4.24 Strata 5.99 Reliability .95
Model, Sample: RMSE .20 Adj (True) S.D. .87 Separation 4.44 Strata 6.25 Reliability .95
Model, Fixed (all same) chi-square: 216.0 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.5 d.f.: 10 significance (probability): .40
Table 1 Examinees Measurement Report(arranged by MN).

The probability curves found show that successful and unsuccessful (2 and 0) facets cross, whereas partials (i.e. 1) are not functioning as a category because the rating is not used quite as often by the raters (13% of times).

The above-illustrated results help to answer the research questions about the effectiveness of the RMA. As this was a pilot study, and there is not a copious body of literature identifying best practices and standard setting methods to establish threshold performances in high stakes, high-level proficiency testing in STANAG 6001-based testing, the main objective and rationale of this conceptual paper was to provide a standard setting method, which would exemplify minimal performance that reflected reading comprehension proficiency in accordance with the STANAG 6001 descriptor. Hence, it was important for the working group members, as representatives of the most well-versed experts in STANAG 6001 Level 4 proficiency, to agree and advance a common understanding and conceptualization of the cut score, and its reflection of the construct.

Discussion of Results

RQ1: *Does the threshold performance, identified through the RMA, accurately reflect the construct of STANAG 6001 Level 4 reading prototype test and validity of uses and interpretations that can be made on the basis of the test scores?*

Both the quantitative analysis using FACETS and the standard setting sessions among the raters show that partial ratings do not contribute significantly to evaluating the responses and determining whether a test taker has fully met the requirements of a Level 4 reader. This could be due to a psychological factor whereby the rater would rather opt for an unsuccessful rating if the examinee's response does not fully meet expectations. STANAG 6001 is a non-compensatory proficiency scale and sustained ability across the three dimensions of content, task and accuracy per level must be demonstrated by a test-taker in order to be awarded that particular level. All raters were well-versed with the STANAG 6001 descriptors, with the test and with the purpose the scores would have in this highly specialized context. However, the non-compensatory nature of the scale may have influenced the raters into awarding more successful and unsuccessful ratings compared to partial ratings. Given the high -stakes nature of this type of context, raters may have preferred to err towards under compensating rather than giving the benefit of the doubt to a response.

RQ2: *How do judges' holistic evaluations correlate with their analytical scores and how reliable are they in predicting the threshold performances*?

As the pre and post ratings show, there is a high inter and intra-rater reliability between initial holistic ratings and final holistic ratings when the cut score identified through the RMA was applied. This denotes consistency within and among the raters, and confirms raters' conceptualization of what it means to be a STANAG 6001 Level 4 reader, based on their experience with the scale, with the test and, mostly, with the inferences and decisions which can potentially be made based on the test scores. Given the high stakes, it is a proven best practice in test development and administration to always have two raters who agree on a rating.

Conclusion

Any approach to setting standards must entail extensive trialling and thorough consequential validity studies to confirm that the minimum cut score established validly reflects the test purpose and use. The RMA approach was applied to scoring test responses of only thirty-three examinees. A major limitation of this study is the sample size which does not allow to make any generalizable inferences of the effectiveness of the method. Further trialling, along with studies of consequential and scoring validity should be conducted to add significant information about

the approach's effectiveness. A statistically significant number of participants, who would be able to pilot the test and the scoring method would be required in order to be able to generalize the appropriateness and applicability of the RMA to official, national testing of Level 4 reading proficiency in accordance with STANAG 6001. Furthermore, additional, external information about the participants' reading proficiency would help to correlate findings with the RMA and determine concurrent validity of the present test design (reading to write).

Finally, further research is needed to comprehend the construct of reading at such high (Distinguished) levels of proficiency, and determine which testing technique best captures such comprehension. Scoring methods reflect the construct and only investigating the true nature of such a construct at this level would be conducive to a valid and reliable cut score setting.

This conceptual paper aims to illustrate a novel approach to setting cut scores which combines the holistic evaluation of expert raters and the analytical ratings of individual constructed responses, called the Retrodictive Modelling Approach on a high-stakes proficiency test. A pattern of minimal combinations of S, U and P ratings, analysed in conjunction with the consistent ratings of initial and final holistic ratings would then yield a cut score. This project shows how important it is to have expert raters on subjectively-scored responses, and that there should always be two raters, if not more, in case of discrepancy in judgment. Rating should initially be done individually and then compared to a second rater's evaluation. Another important finding is the fact that analytical scoring contributes to guide the rater as to whether the final rating should be changed from the initial rating. A holistic rating alone does not seem to consistently predict the performance of the sample.

This project is far from over. Consequential validity studies which investigate the consistent and meaningful uses of the cut score found through the RMA needs further research within and across NATO member and partner neations which have adopted the STANAG 6001 proficency scale and need to test Level 4 proficiency.

Acknowledgements

This project would not have been possible without the tireless support of Ms. Peggy Garza (U.S. Marshall Center, Germany) and Mr. Keith Wert (U.S. Marshall Center, Germany), BILC secretary and Chair at the time of this project. Additionally, special recognition must be given to the Working Group members who contributed to the project as either test developers or raters, and who devoted countless hours to this work: Ms Petya Georgieva (National Defense College, BLG), Ms Nancy Powers & Dr. May Tan (Military Personnel Generation Group, St., CAN), Mr. Käre Kildevang and Dr. Allen Christiansen (Royal Danish Defense College, DNK), Dr. Donald Sturges (Bundessprachennamt, GER), Mr. Gerard Seinhorst (National Defense Language Center, NL), Ms Annette Nolan & Mr. Keith Farr (Swedish De-

fense College, SWE), Mr. David Oglesby (U.S. Marshall Center, Germany), and Dr. Martha Herzog (DLIFLC, USA).

Finally, the authors wish to thank Dr. Troy Cox (BYU, USA) for his expertise and help along with Dr. Ray Clifford (BYU, USA) for his contribution to understanding Level 4 proficiency according to STANAG 6001.

References

Alschuler, C. & Moussa, N. (2002). *Testing reading at ILR Levels 4, 4+ and 5: A Tester Training Model*. Presentation to the Interagency Language Roundtable Committee, Foreign Service Institute, Washington, D.C.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In Thorndike, R.L. (Ed.), *Educational*. *Measurement* (pp. 508-600). American Council on Education.

Language Needs Analysis. (2015). Unpublished NATO BILC internal document.

Chang, L., Dziuban, C., Hynes,M., & Olson, A. (1996). Does a standard reflect minimal competency of examinees or judge competency?" *Applied Measurement in Education 9 (2)*: 161.

Child, J.R. (1998) Language skill levels, textual modes and the rating process. *Foreign Language Annals 3*: 381–397.

Cizek, G., & Bunch, M. (2007). *Standard- setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications.

Davis, A. (2003). Three heresies of language testing research. *Language Testing 20(4)*: 355– 368. https:// 10.1191/0265532203lt263oa.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance processes. *Applied Measurement in Education 10*: 61-82. https://doi:10.1207/s15324818ame1001_4.

Edwards, A. L. (1996). Reading proficiency assessment and the ILR/American Council on the teaching of foreign languages text typology: A reevaluation. *Modern Language Journal 80*: 350–361.

Kane, M. T. (1998). Criterion bias in examinee-centered standard setting: Some thought experiments. *Educational Measurement, Issues and Practice 17*: 23–30.

Lowe, P. (1998). Keeping the optic constant: A framework of principles for writing and specifying the AEI definitions of language abilities. *Foreign Language Annals, 31(3)*: 358–380.

Shrock, S. & Coscarelli, W. (2007). *Criterion-referenced test development. Technical and legal guidelines for corporate training (3rd edition)*. Internet: Pfeiffer & Company.

Shulruf, B., Poole, P., & Wilkinson, T. (2015), The Objective Borderline Method: A probabilistic method for standard setting, *Assessment & Evaluation in Higher Education 40(3)*: 420–438. https://dx.doi.org/10.1080/02602938.2014.918088.

Shin, S., & Lidster, R.(2017), Evaluating different standard-setting methods in an ESL placement testing context. *Language Testing 34(3)*: 357–381. https://DOI:10.1177/026553226646605.

Standardizing Agreement 6001 Proficiency Scale (STANAG). (2015). Unpublished BILC document. www.bilc.org.

Re-Conceptualizing Language Programs to Achieve Level 4

Christine Campbell, Campbell Language Consultats (USA)

Interagency Language Roundtable (ILR) Language Skill Level Descriptions, which delineate levels of proficiency using a scale from 0 through 5, can portray the individual at Level 4[43]. In the U.S., this level is sometimes referred to as "The Diplomat" by American Council on the Teaching of Foreign Languages (ACTFL, ILR Workshop, 2018): that is, one with access to a vast linguistic repertoire. Within the context of professional needs, such a person can understand all forms and styles of speech, read fluently and accurately all styles and forms of the language, and speak fluently and accurately on all levels. Scholars and practitioners who have written about learners either approaching or at this level highlight certain shared characteristics while recognizing their uniqueness (Davidson & Garas, 2018; Davidson & Shaw, 2019; Ehrman, 2002; Farraj, 2006; Ingold, 2002; Leaver, 2003; Leaver & Atwell, 2002; Leaver & Campbell, 2015, 2020; Leaver & Shekhtman, 2002; Robin, 2008; Soudakoff, 2004).

Recently, Franke (2020) conducted a qualitative case study that explored how persistence, study abroad, motivation, and learner autonomy play into the pursuit of Distinguished speaking proficiency in particular. The data analysis of interviews with the non-native speakers revealed that attaining Level 4 "was a highly personal pursuit, characterized by different motivations based on the choice of a foreign language, engagement in the target culture, grit, and time. Overall, the participants

43 Level 4 in the ILR scale is equivalent to the Distinguished Level on the ACTFL Proficiency Guidelines scale. For more information, see www.govtilr.org and www.actfl. org/publications/guidelines-and-manuals/actfl. In the article, level references will be to both the ILR and the ACTFL scales at the start, then only to the former (cf. ILR/ACTFL).

were highly self-efficacious learners, many married to foreign-speaking spouses, and spent [sic] extended periods in the foreign culture and community" (p. iv).

Corin (2020) proposes that if learners are to have a realistic chance of reaching Level 4, the second language (L2) learning process must be designed in such a way that they can: 1) approach Level 3 early enough in the course of learning (typically, in their organized courses of study) to "enable an ascent" to the mountain summit--near-native proficiency; and 2) arrive at Level 3 already possessing specific "equipment," here understood as "expanded learning capacity," needed for the final scale to the top (pp. 1 and 25, respectively). The "capacity" is comprised of skills such as "strategic ambidexterity," including both preferred [learning] strategies [e.g., metacognitive, cognitive, social, affective, and communicative as per Shekhtman, 2003b] and those typical of learners with opposite 'native' learning styles, expanded sociocultural awareness, competencies and permeability" (p. 25).

This chapter will focus on a separate aspect of the Level 4 dialogue—concrete, proven measures that Level 4 language programs can take when re-conceptualizing. Specifically, it will describe actions implemented by the Directorate of Continuing Education (CE), Defense Language Institute Foreign Language Center (DLIFLC), from 2006 through 2016 to facilitate learners attaining high levels of proficiency in standardized exit tests—the *Defense Language Proficiency Test* (*DLPT*), for Listening Comprehension (LC) and Reading Comprehension (RC) and the *Oral Proficiency Interview* (*OPI*), for Speaking (SP). At the time, CE was responsible for teaching three types of courses, among others—Intermediate, Advanced, and Defense Threat Reduction Agency (DTRA) Interpreting[44].

The far-reaching goals to be Distinguished from graduation requirements of the Advanced and DTRA courses were Level 3+ and above in LC and RC and Level 3 and above in SP. Graduation requirements for the two courses were, and are still today, considerably less demanding: Advanced Course—Level 3 in LC and RC and Level 2 in Speaking (SP); DTRA—Level 2+ in LC and RC and Level 2 in SP. The entry requirement for the Advanced Course is Level 2+ in LC and RC; DTRA, Level 2 in LC and RC. There is no entry requirement in SP. The length of the Advanced Course is 19 weeks; DTRA, 47 weeks. Given the surprisingly low entry and graduation requirements, the consistently high scores historically achieved in the CE Program, the DTRA Course in particular, are especially noteworthy.

First, this article will review the institutional context of the CE language program. Then, it will study measures put in place to bring about fundamental change in curricular design, teacher education, and learner education (here understood as instruction in learner styles and strategies, orientations about pedagogical approaches and classroom protocol, etc.) as part of CE's quest to reach ever higher levels of proficiency. Finally, it will examine the DTRA Interpreting Course Program, where learners regularly attain Level 3+ and above in at least one of the three

44 The article narrative is set in the past given it relates to activities in CE from 2006-2016.

skills. At this point in 2020, for example, 63% of DTRA learners have graduated at Levels 3+ and 4 in at least one of the three skills[45].

The Institutional Context of the CE Language Program

DLIFLC is the premier language learning facility of the United States government, where close to 4,500 learners, the majority potential military linguists, study one of 17 languages. Its faculty of 1800 is comprised of teachers, curriculum developers, teacher educators, test developers, technology specialists, and more. Due to geopolitical security concerns facing the United States, DLIFLC was urged by the government in 2012 to re-look its *modus operandi* as part of an Institute-wide initiative designed to increase by 30% to 80%, depending on the language program, the number of graduating learners from the Basic Course (BC) at Level 2+ in LC and RC and Level 2 in SP.

The in-depth needs analysis conducted at the outset of the initiative ultimately led to a call for a different educational approach. To reach higher levels of proficiency, the approach would have to go beyond Communicative Language Teaching (CLT). Leaver, who became Provost of DLIFLC in 2013, had been applying principles of Transformative Learning (TL) which identifies "perspective transformation" as the central learning process, into her teaching since the 1990s (Mezirow, 2000, p. xi). Seeing the potential of TL for language learning, where learners are exposed to different perspectives on the realities of life as part of their language and culture studies, she disseminated its principles at the Institute, where they inform practices until today.

TL, according to its discoverer, Mezirow (2000), is "the process by which we transform our taken-for-granted frames of reference (meaning perspectives, habits of mind, mind-sets) to make them more inclusive, discriminating, open, emotionally capable of change, and reflective so that they may generate beliefs and opinions that will prove more true or justified to guide action…Frames of reference are the results of ways of interpreting experience" (pp. 7-8 and 16, respectively). He elaborates: "[W]e transform frames of reference—our own and those of others—by becoming critically reflective of their assumptions and aware of their context—the source, nature, and consequences of taken-for-granted beliefs" (p. 19). Individuals undergoing transformation pass through phases in which they "clarify meaning" as they experience the world (p. 22). The first phase—often a "disorienting dilemma," provokes self-examination and critical assessment of assumptions that lead to the exploration and adoption of new roles, relationships, and actions, which culminate in a "new perspective" (p. 22).

Widely discussed and applied in the field of adult education for over years, TL had received limited attention in world language education (e.g., Arce, 2000; Foster,

45 More specific information is inaccessible for security reasons.

1997; Goulah, 2006; Leaver & Granoien, 2000; McClinton, 2005; Osterling & Webb, 2009) until more recently. In 2006, Leaver introduced select principles at CE; in 2013, she incorporated them as part of the Institute initiative to raise proficiency levels. Its formal adaptation to language learning as "Transformative Language Learning and Teaching" (TLLT) is the result of a collaborative effort by professionals such as Leaver, Davidson, and Campbell (2020) and Nyikos and Oxford (Leaver, Campbell, Nyikos, & Oxford, R., 2019).

The primary goals of TLLT are personal transformation that results in bilingual and bicultural competence and learner autonomy. Common features of TLLT as currently being used in diverse learning contexts include the following:

- "Materials and daily communications are authentic and unadapted, beginning at the earliest levels of instruction.
- The classroom is immersive. Immersion in-country reflects the typical life of the native speaker of the same age to the extent practicable.
- Personal transformation involves cognitive, emotional, and cultural shifts occurring within the individual: that is, developing self-awareness, resolving disorienting dilemmas, identifying cognitive distortions, managing emotions, and integrating two (or more) cultures on their own terms.
- Highly individualized programs are informed by learning styles and strategies and the "invisible classroom" emanating from inherent personality variables.
- Open architecture curricular design (OACD), a term taken from a parallel concept in computer design that allows for interchangeable parts, supports increasingly textbook-free classrooms as learners develop greater proficiency and teachers modify syllabi corresponding to learners' changing needs.
- The grading system uses formative assessments and feedback, with occasional summative assessments (projects, presentations, contracted assignments, and portfolios), that integrate outcome and process, instead of separating them.
- Programs empower learners to take charge of their own learning.
- Program design and supervisors empower teachers to take charge of their own classrooms as advisors, mentors, coaches, planners, and strategists (Leaver et al., 2020).

TLLT principles operate within a context that is grounded in the national standards (The National Standards Collaborative Board, 2019) and is content-based and learner-centered.

Contrasting TL with TLLT, Leaver posits: "At least for now, these types of TLLT activities differ in important ways from contemporary aspects of transformative learning in adult education: whereas transformative adult education often focuses on creating change in society, TLLT seeks to understand and create a synthesis with the other society" (p. 19, Leaver, 2020).

Due to the history of high scores attained by CE learners upon graduation, senior management at DLIFLC analyzed the CE program closely for practices that would transfer to the Basic Course. Although the learners were different in that those in CE were professional military linguists in their early and mid-20s and those in the BC were budding military linguists aged 18-21, the over-arching approach toward curricular design, teacher education, and learner education was deemed applicable. A description of the three areas follows.

Curricular Design

Many educators today who keep abreast of the latest developments in curricular design practice the "backward-design" principle of modern curricula, which Wiggins and McTighe (2005) posit starts with the performance goals and works backwards to the performance assessments,[46] whether formative or summative. Sandrock (2010) distills the principles of backward design like so: "[I]dentify clear performance goals, then create assessment tasks through which students will demonstrate the performance goals, and finally plan what students need to know and be able to do in order to be successful in the assessments—that is, the vocabulary, grammatical items, and language functions" (p. 170).

The backward-design principle is one aspect of Open Architecture Curricular Design (OACD), a fundamental feature of TLLT that was first introduced to CE faculty by Leaver in 2006. OACD is a flexible framework that encourages teacher–learner negotiation through the use of a theme-based syllabus (versus a textbook), generally beginning at Levels 1+/Intermediate High and above though there is evidence of the successful introduction of OACD at much lower levels (Clifford, 1988; Duri, 1992), depending on factors such as teacher and learner readiness. A fundamental principle of TLLT, OACD has been a critical factor in CE's academic successes.

46 In the language learning context, Sandrock (2010) confirms that performance assessments "provide a realistic description of expected student progress in developing proficiency in the language" (p. 171).

Below, measures taken to reconceptualize curricular design in CE:

Adopt OACD

Leaver, who had implemented aspects of OACD without using the term as early as the 1980s at the Foreign Service Institute (Leaver, 1989), shared it with faculty at CE when she assumed the position of Associate Provost in 2006. In OACD, a textbook orients teachers through Level 1/Intermediate Low and Intermediate Mid to facilitate standardization in the development of structural and lexical control. Generally beginning at Level 1+/Intermediate High and above, but occasionally at lower levels, depending on teacher and learner readiness, there is no textbook but rather a theme-based syllabus guiding teachers. Common features exhibited by OACD teachers are:

- use of authentic materials from day one;
- deliberate, continual use of the target language;
- learner delivery of content [From the start of instruction, there is ongoing learner involvement in both the delivery and selection of content];
- project/scenario-based instruction;
- development and use of higher-order thinking skills;
- use of formative assessments;
- integration of both formal and colloquial language;
- integration of non-standard language;
- incorporation of collaborative learning, such as group presentations and projects based on learner research;
- use of a wider variety of listening and reading genres across the full spectrum of social media platforms, such as Instagram, Twitter, WhatsApp, possibly Facebook, LinkedIn, blogs, and so on (Campbell & Sarac, 2017);
- systematic defossilization;
- focus on stylistics, including use of register;
- focus on discourse analysis;
- incorporation of super-authentic (Cohen, 2015) language – language spoken by two or more people with ambient noise, grammatical mistakes, fillers, and so on; and

- top-down and bottom-up processing of high-level presentations on topical domains such as politics, economics, and history by guest speakers (Campbell, 2020).

Redesign the Assessments

As Dabaneh and Yuan (2020) observe, "[o]ver the past four decades, testing specialists have pondered the role of testing and to what extent it can be "educative," informing learners about their strengths and weaknesses and preparing them for the next level of learning (Wiggins, 1998, cover)" (p. 1). The advent of task-based instruction brought about the deliberate integration of assessment and instruction, which in turn led to [instruction] "embedded assessment," defined in 2003 by Spence-Brown as "the use of tasks which serve a pedagogical purpose for assessment" and has the advantage of providing robust feedback and allowing for the washback effect to motivate students to achieve specific short term goals"(p. 36). According to this researcher, the problem with this type of "formative" assessment is that the product of learning "does not represent what a student can do unassisted, after the learning cycle is completed, but what they can do with support during the learning process" (p.36).

Also in 2003, Poehner and Lantolf, expanding on previous work, published "Dynamic Assessment of L2 Development: Bringing the Past into the Future," where they examine "Dynamic Assessment" (DA) as a developmental approach to assessment and instruction derived from Vygotsky's theory of the Zone of Proximal Development (ZPD). They define DA "as a procedure whose outcome takes into account the results of intervention. In the intervention, the examiner teaches the examinee how to perform better on individual items or on the test as a whole. The final score may be a learning score representing the difference between pretest (before learning) and posttest (after learning) scores, or it may be the score on the posttest considered alone" (pp. 1-2). In 2017, Poehner, Davin, and Lantolf reiterate that "[d]ynamic assessment, or DA, departs from the traditional distinction between formative and summative assessment, as it understands teaching to be an inherent part of all assessment regardless of purpose or context" (p. 243).

Outside of the DA context, the distinction between formative and summative assessments is generally observed. Perhaps Bachman's (1990) definition, cited by Poehner and Lantolf in 2005, is the most succinct: "In the language testing literature, FA [Formative Assessment] is usually contrasted with Summative Assessment on the grounds that the former is intended to feed back into the teaching and learning process while the latter reports on the outcomes of learning" (pp. 60–61)" (p. 233).

When redesigning assessments to focus on teaching versus testing, for example, it is critical to understand fully the differences between formative and summative

assessments or a language program can perpetuate the over-testing trend found in many learning institutions. In *The Keys to Assessing Language Performance: A Teacher's Manual for Measuring Student Progress*, Sandrock (2010) provides a practitioner-oriented definition of the two, with helpful examples of each: "Formative assessment ranges from quick learning checks to activities guiding students to more independent use of language.…In summative assessment, students demonstrate to themselves and their teacher that they can apply the lessons learned, the skills acquired, and the knowledge gained in the unit of instruction. This is when students produce language on their own and show what they are able to do as a result of the instruction. Summative assessment is a new application of the individual elements of vocabulary and grammar assessed at the formative level. Through summative assessment, students showcase the level of proficiency acquired" (pp. 62-64).

Balancing the two types of assessments, teachers are responsible for ensuring feedback is continually provided to learners. For Sandrock, feedback in formative assessment is "specific and highly focused, as students are learning and practicing various building blocks in preparation for the final unit level performances (e.g., commenting on pronunciation, use of a specific structure, or ability to elaborate and provide more detail). In summative assessment, feedback is more broad and holistic, where the teacher steps back and looks at the overall performance (e.g., commenting on student ability to get meaning across, maintain a conversation, or organize a strong argument)" (pp. 62-63).

Sandrock provides helpful examples of formative vs. summative assessments. Below, one of each:

Level: Level 0+, Emerging Level 1/Novice
Topic: What Makes a Good Friend? How Am I a Good Friend to Others?

"Formative Tasks:

- Ask other students what they like to do.
- Ask other students what they do to help others.
- In small groups, list characteristics of a good friend.
- Hear statements and identify if a good friend would or would not do that.
- List what friends do in school, what friends do outside of school.
- Write down two things you will do to be a better friend to others.

Summative Tasks:

- Listen to a conversation and on a list of things that good friends do, check off what these friends say they do for each other.
- Identify with a partner what good friends do and don't do.

- Write a letter to a friend identifying three things you plan to do as a good friend and three things you will not do as a good friend" (pp. 61-62).

Sandrock also highlights the role of rubrics when distinguishing formative from summative assessments like so: "Formative assessments, being simpler, shorter, and more limited in scope, may not necessarily require a rubric. Summative assessments elicit a more complex performance from students and therefore warrant the time needed to develop a rubric that can provide extensive feedback" (p. 88).

Informed by the developments in the assessment field like those reviewed above, in the mid 2000s CE revamped the entire assessment program to focus on formative assessments, with occasional summative ones. The re-conceptualizing led to the situation_Sandrock describes in the conclusion to *Keys*: "When student progress is measured through performance assessment and effective feedback, students know what they can do in the new language they are acquiring and what they need to do to improve their proficiency and increase their confidence in using the three modes of communication. This is the road map to guide language teaching and learning" (pp. 190-191).

Teacher Education

Teacher education[47] is often recognized as perhaps the most significant factor in improving learner performance. Well-designed, task-based, content-based, learner-centered instruction informed by the world -readiness standards sets up optimal conditions for learning which learners can use to further their goals as autonomous persons. Below, measures taken to reconceptualize the CE Teacher Education Program:

Foster a Community of Understanding and Practice

Just as it is critical for learners and teachers in today's learning space to work together to create a learning community where both groups share an understanding of expectations and strive to reach the same goals, so it is crucial to forge the same kind community of understanding and practice among key participants in an educational setting (Bailey & Freeman, 1990). The community of participants, which includes learners, faculty, and administrators, can benefit from adopting a team approach to dealing with the challenges of a language program, whether an increase in learners not completing graduation requirements, an outdated or rigid curriculum, understaffing, a percentage of the faculty who is pedagogically unpre-

47 Teacher education is interpreted broadly here to encompass both current classroom teachers and those who have joined the administration such as Deans, Department Chairs, and Team Leaders. (At DLIFLC, a teaching team model where two to six teachers collaborate daily on the creation [in some cases, only adaptation] of the curriculum, is used.)

pared, and more. The team approach is especially effective during times of change as it diffuses the anxiety often felt by participants.

In the case of CE, when Leaver brought aspects of TLLT such as OACD to the Directorate in 2006, the faculty were compelled to learn a new way of creating curricula and teaching. Understandably, changes in faculty performance standards due to the changes in pedagogical orientation caused anxiety. The team approach facilitated the creation of a community of participants who rallied to meet the new requirements. The premise of the approach is that all participants share in program successes and failures. Successes were celebrated; failures were opportunities to learn. When, for example, learners did not meet the minimum graduation requirements in standardized exit tests, the administration first recognized the efforts of the conscientious faculty who had put forth every effort to succeed. Then, the administration and faculty held meetings to discuss and document what worked and what did not. Learners, accustomed to voicing their perspectives with the administration and faculty in periodic "Sensing Sessions," which are described in detail in the "Learner Education" section below, shared their comments in focus groups. The information gathered at the meetings and focus groups was included in a report that acted as a guide for future courses so course strengths could be enhanced and weaknesses eliminated.

The team approach was also used when crafting the in-service Teacher Education Program. Several times per month, two to three faculty made presentations about relevant language learning issues as part of the Professional Development Series held after the class day from 3:30-4:30 p.m. The Series was organized by faculty for faculty.

Set High Expectations

Goethe, the 18th century author, stated about expectations: "Treat people as if they were what they ought to be, and you help them to become what they are capable of being" (https://quoteinvestigator.com/2018/10/09/capable/). The CE community of participants has consistently confronted challenges, such as the setting of higher expectations, with optimism and fortitude, knowing they are supported by an organizational culture that promotes teamwork. For example, when new performance standards were issued in CE in 2007 due to the changes in pedagogical approach made by Leaver, the community adapted to the higher expectations, understanding that each community member would accept a greater degree of responsibility. They reacted similarly when the new graduation goals in the BC—80% 2+ in LC and RC by 2021—were announced by DLIFLC in 2013. Although CE does not teach the BC, the higher percentages of 2+/2+ meant that CE had to graduate

learners in the Intermediate Course at higher levels, and, as a consequence, the Advanced Course took on even higher goals. Today, the CE community continues to display the same flexibility and dedication.

Redesign the Teacher Observation Program

Regarding teacher observations, the Institute requires Chairs to conduct several summative ones, where information gathered is used to inform the annual teacher performance evaluation. CE re-designed the teacher observation system with a focus on the formative value of observations, converting a typically anxiety-provoking event into an opportunity to re-visit one's strengths and learn how to improve areas of concern. Chairs announced the formative observations, emphasizing that information collected would be used solely as a reflection tool for teachers. Faculty members were also urged to conduct peer-to-peer teacher observations.

Promote Access to Institute-Wide Professional Development Opportunities

Through today, the CE administration ensures faculty can avail themselves of the wide array of teacher education opportunities at DLIFLC that are critical to the Institute's initiative to increase graduation requirements. Below, a listing:

- Faculty, Team Leaders, and Administration:
  - Advanced Language Academies. The academies, which are open to other US government language schools and Foreign Language Flagship Programs, afford academic leaders, academic specialists, teacher educators, and faculty the opportunity to discuss the theory and practice of TLLT. Typical topics were Overview of SLA Theories; OACD; Task-Based Instruction; Content-Based Instruction; Learner-Centered Instruction; Formative and Summative Assessments; Genre and Authentic Materials.
  - Summits. Summits focused on a particular topic, such as "Actualizing Open Architecture Curricular Design" (May, 2016), are organized several times per year.
  - 1-Day Language Learning Conferences. Twice yearly, the Faculty Development Division of DLIFLC holds one-day language learning conferences with 25 faculty presenting on topics of interest to professionals. The process for selecting presenters includes blind-review of the proposals.
- Teachers: Teacher Education Series are routinely organized by the Provost Office, e.g., "Reaching 2+: Sharing Successes" occurred every

month for one year. At these sessions, faculty from across DLIFLC described the record-breaking results achieved in their courses.

- Team Leaders: Workshops aimed at the approximately 75 Team Leaders on the main campus are held about subjects such as "Implementating OACD."
- Administration: Workshops take place, e.g., four sessions where participants co-created and presented on "Actions Plans for Reaching Higher Levels of Proficiency."

Learner Education

Learners accustomed to traditional, teacher-centered language learning contexts can be disconcerted when first presented with learner-centered instruction because they have limited experience taking responsibility for their learning, practicing autonomy while accessing learning opportunities 24/7, engaging in collaborative learning situations where they work in pairs and group work, and more. CE teachers offer learners an orientation that helps learners with the transition from teacher-centered to learner-centered instruction. Language learning style and strategy instruction for learners, which is fundamental for making the critical leap from Level 3/Professional to Level 4/Distinguished, follows the orientation. Below, measures taken to reconceptualize learner education in CE:

Provide Learning Style and Strategy Instruction

CE faculty introduce learners to learning style and strategy instruction when they start their courses. Learners are administered a battery of instruments designed to provide information on learning style, which Leaver posits is the "most significant" individual difference related to language learning (2003, p. 53). This battery consists of (1) *E&L Cognitive Styles Construct* (Ehrman & Leaver, 1997, 2002, 2003) for cognitive style; (2) *Myers-Briggs Type Indicator* (*MBTI*) (Myers & Myers, 1980) for personality type; and *Barsch Learning Styles Preference Form* (Barsch, 2003) for sensory preferences. Data gathered from the different instruments is included in an Individualized Study Plan (ISP), providing insights into learner style and which learning strategies can be especially beneficial.

When conducting learning strategy instruction, Oxford's (1990, 2016) classic taxonomy of metacognitive, cognitive, social, and affective learning strategies is the preferred guide. Shekhtman (2003a, 2003b, 2003c, Shekhtman et al, 2002) adds communication strategies as a category, asserting they are especially important at the higher levels.

Leaver and Shekhtman (2002) posit that the nature of learning strategies necessarily changes according to proficiency level. For example, strategies for comprehending authentic printed texts such as guessing from context, using titles and visuals appearing with the RC text or LC passage as clues, circumlocution, use of cognates clues, and such are helpful at the lower levels. Risk-taking applies to all levels. Below, some examples of learning strategies for high-level learners recommended by Leaver (2003):

- Metacognitive: Planning study activities; practicing advance organization; observing learning effectiveness; evaluating progress; rewarding progress; and revising plans.
- Cognitive: Memorizing/associating; compensating (rarely needed); processing (recycling, manipulation for practice, use of synonymy, and "-lect" switching, which is understanding the differences among sociolects, dialects, and idiolects).
- Social: Asking a native to proof written text or listen to a speech; reviewing movies and books with native speaker friends and colleagues; asking a native-speaker mentor to help with the development of cognitive strategies.
- Affective: Practicing positive self-talk after hearing criticism from native speakers about a linguistic mistake or socio-linguistic misstep; keeping a journal.
- Communication: for LC—Using the interlocutor, managing the input; for RC—comparing genres, using topic saturation, and comparing interpretations; For SP—developing and using automatic discourse, using authentic materials, using conversational repair, embellishing speech, and controlling the conversation; Writing—developing and using automatic discourse, using authentic materials, using process writing, embellishing, and obtaining authentic feedback.

Use Individualized Study Plans (ISP)

An ISP referred to briefly earlier, is a key tool in the arsenal of the learner dedicated to reaching Levels 3+ and beyond. According to Leaver (2003b), the purpose of an ISP is "to assist students in organizing their short-term and long-term learning goals and activities" (p. v). The *ISP*, which all learners in CE work with their teachers to construct, reflects the learning objectives, learning experiences, interests, learning style, and financial/time constraints of the learner. Although no specific format is required, most *ISPs* include "courses; study, work and travel abroad; independent study; reading; use of the Internet; work with a native speaker; development of friendships with speakers of the language; writing to pen-pals (and/

or friends and relatives); practica and internships; watching television; listening to the radio and tapes; becoming acquainted with music, arts, and dance; foreign assignments; and periodical assessment of progress" (Leaver, 2003b, p. 4). Included is important information about learning style and learning strategies, which were discussed in the preceding section.

The data can be organized accordingly:

- Skill-based plans (LC, RC, SP, and Writing [WR]);
- topic-based plans (specialized and general, based on, for example, work needs or future plans for the language);
- proficiency-based plans (development of discourse, linguistic, sociolinguistic, strategic, and socio-cultural competence);
- and chronological plans.

Set up a Diagnostic and Dynamic Assessment Program

Dynamic Assessment (DA), which was briefly examined in the "Curricular Design" section above, and Diagnostic Assessment[48] are both valuable learning tools focused on learner improvement. Currently, DLIFLC is implementing both. Diagnostic Assessment is a formative learning tool created at DLIFLC in 1996 to diagnose learner strengths and weaknesses in the four skills of LC, RC, SP, and Writing (WR) (Steven Koppany, cited in Cohen, 2003). It both identifies problem areas in the four skills and provides clear recommendations for how the learner can improve performance. Linguistic breakdown is a focal point because it provides critical information to the assessor about learner problem areas.

Diagnostic Assessment consists of a three-skill interview in LC, RC, and SP with two assessors and a learner that lasts 90 minutes, although it can be extended according to learner needs. Thirty minutes are allotted to assess each skill. Diagnostic Assessment begins with the SP component—an oral interview based on the ILR OPI that establishes the learner's current proficiency level in SP and serves to familiarize the learner with the process. One assessor engages with the learner, clarifying that Diagnostic Assessment is not a test but a way to help the learner remedy problem areas. The other assessor takes notes about learner performance. Concrete recommendations for dealing with the problem areas are prepared by the assessor and delivered to the learner in a *Learning Plan*, which complements the *ISP* described earlier.

In the RC portion, the learner is presented with at least three texts—one at the learner's proficiency level, as indicated by performance in the SP part, one at a bit higher level, and one at a much higher level. The learner reads silently and then dis-

48 Given the acronym for Diagnostic Assessment is identical to that of Dynamic Assessment, the former will be referred to by its full name.

cusses its main idea and details, in either the target language (TL) or English, with the assessor. Given the goal is to assess RC, not SP, use of the TL is not required although learners at Level 3 and above generally prefer the TL to English. The assessor asks the learner pointed questions to determine levels of understanding.

The LC part can assess either interactive and non-interactive LC or both, depending on institutional needs. In the former, one assessor simultaneously evaluates SP and LC. Cohen (2003) suggests an alternative—lengthen the interview to focus on SP in the first part and LC in the second part. Concerning non-interactive LC, the learner is asked to listen to passages, which can be replayed so memory is not a confounding variable. As with RC, given the goal is to assess LC, not SP, use of the TL is not required, although learners at Level 3 and above generally prefer the TL to English. The assessor asks the learner pointed questions to determine levels of understanding.

Cohen (2003) asserts that "[a] Diagnostic Assessment is especially advantageous for higher level language learners" because "[e]specially at L3, students do not know what their gaps are" (p. 29).

Promote Learner Voice

Learners grasped soon after arriving in CE that they were an essential part of the community of participants cited earlier. The OACD principle—Ongoing learner involvement in the selection and delivery of content, requires learners lean forward and take responsibility for their learning throughout the course. Sensing Sessions are opportunities for learners to provide weekly, monthly, bi-monthly or quarterly feedback to a teacher(s) and/or administrator(s) about aspects of the academic program such as the teachers, curriculum, use of technology and more. The feedback, which is uncensored but respectful in tone, is an important source of information to teacher(s) and/or administrator(s) which can be used to improve processes. The person in charge of the Sensing Session, who can be the teacher of the learners, another teacher, a department chairperson, a Dean, etc., arrives with a laptop or flipchart to take notes. The person alerts the learners at the start that it is critical to (a) be fair and honest; (b) be respectful when providing constructive criticism; (c) not project personal frustrations unrelated to learning against the teacher(s) and/or administrator(s); and (d) be cognizant of the power they have when asked for their opinion in this type of forum. The learners are told they can comment as often as they wish, first citing the positive aspects, then the areas of concern. At the end, the person in charge asks the learners to vote in order to identify what areas of concern might be shared by the majority. These areas are then communicated to the other teachers in the case of team teaching. The feedback to the teacher(s) is formative, not summative, i.e. the feedback is not used by the administrator(s) to evaluate the teacher(s); rather it is used as a tool for professional development. The teacher or

teaching team then systematically works to correct the areas of concern. Sensing Sessions are an efficient way to enhance inter- and intra-group communication.

Defense Threat Reduction Agency Program

The Defense Threat Reduction Agency (DTRA) Interpreting Course in Russian at DLIFLC is an intensive program for learners of Russian whose assignments within the government require a high level of language proficiency for conducting weapons control work in Russia. Thirteen faculty are dedicated to achieving the goals of Levels 3+ and above in LC and RC, and Level 3 and above in SP. Starting in February 2020, DTRA learners who meet graduation requirements will receive 87 college credits, the majority 300 to 400 level courses, as per the American Council on Education (ACE).

When creating the curriculum, teachers apply principles of OACD. The theme-based syllabi covering 24 general topic areas such as history, economics, political science, technology, and health, and specific topic areas such as the latest developments in foreign policy is negotiated with learners. Teachers, working with learners, create daily assignments based on current authentic materials that are previewed before class, following the flipped classroom approach. Teachers also develop new courses based on trend analyses across and within groups of graduating learners. For example, to help learners develop greater register range, DTRA teachers designed a Stylistics Course that highlights the distinctive styles found in various literary and nonliterary genres and in the works of individual writers. Below, the themes examined in the Stylistics Course:

The notion of norm and variability in language use. Standard and colloquial styles. Word order and sentence structure: comparative typology of the Russian and English languages.
Classification of functional styles of the Russian language. Description of linguistic means, sub-styles and speech genres typical of each style in the Russian language.
Emotional and expressive aspects of the language. Expressive suffixes in the Russian language. Tropes: similes, types of metaphors, irony, hyperboles, litotes.
Types of phraseological units: set phrases, idioms, proverbs and sayings, their role in stylistics. Challenges for interpreters.

Teachers, acting as mentors/coaches/advisors, continually urge learners to use higher-order thinking skills as they research topics for roundtables and presentations. A formal roundtable discussion on a controversial topic takes place every two weeks. Learners study the topic in depth to be ready to discuss it in detail. One

of the learners acts as moderator and organizes the discussion by preparing three to four thought-provoking questions.

In addition to roundtable discussions, learners also make frequent 10–15-minute presentations on topics they have researched. The presentations are followed by elaboration activities such as debates, which encourage analysis, synthesis, evaluation (and later, creativity) as per Bloom's (1956) taxonomy of learning objectives. Learners have ample time to prepare for debates, working collaboratively in teams. One instructor works with each team separately. All instructors act as judges, although in one debate more advanced learners in the same program are judges. Teachers direct learners to relevant websites, provide higher-level, register-appropriate expressions and connectors, and review learners' work before it is presented. To develop greater accuracy in the target language, learners attend required grammar sessions and individualized speaking sessions with teachers, where they review the fossilized errors detected by teachers during the activities.

In the DTRA course, learners take interpreting excursions to local sites, though this activity has been curtailed because of the pandemic. Examples of sites are: Community Hospital of the Monterey Peninsula, the USS Hornet, the local Police Station, US Coast Guard Station, Colton Hall Museum, Monterey Bay Aquarium, Monterey Airport, and Naval Postgraduate School Laboratories. Pairs in the class of six prepare a 10-15 mt. presentation in Russian about some aspect of the site to be visited. Learner A gives the presentation; Learner B engages in simultaneous, consecutive interpreting into English. Learners have two weeks to prepare for the interpreting excursion. The real-life activities help prepare learners for their future work. Due to the number of interpreting excursions from eight to 16.

Another activity is the in-class mini-excursion where Learner A gives a detailed briefing to peers in Russian while Learner B interprets into English (and vice versa). Both briefers and the audience are expected to ask questions, making the experience interactive. Interpreting for guest speakers, who make interactive presentations lasting up to two hours, is a challenge for learners who are about to graduate.

In addition to the activities listed, DTRA learners engage in two-hour negotiation scenarios, where they are immersed in a variety of contexts, e.g., discussion between union officials and company representatives about salary and benefits, etc. Learners receive critiques on their performance from both teachers and peers.

Over the years, the results in LC, RC, and SP in the DTRA Interpreting Courses have continually risen due, in great part, to the program improvements outlined earlier. In 2017, 35% achieved Levels 3+ and 4 in at least one of the three skills; 2018, 46%; 2019, 49%; 2020, 63%. (More specific information is inaccessible.)

Conclusion

Reconceptualizing language programs with the goal of Level 3+ and above is a complex enterprise that presents challenges in the areas of curricular design, teach-

er education, and learner education. Examining frameworks that have consistently produced high levels of proficiency can provide ideas for improving learner performance. While some might question the generalizability of some of the principles and activities in the CE framework to the K-16 environment, the basic concepts are directly applicable to learners aged 16 and above.

While every learning context is *sui generis*, a veritable universe onto itself, given the interplay of myriad variables in the areas of human resources, materials, and fiscal resources, measures that apply to most contexts do exist that can be systematically applied to increase the probability for success. This chapter has presented a sampling of measures for the purposes of reflection and possible adoption.

References

American Council on the Teaching of Foreign Languages. (2018). *Testing Workshop*. Interagency Language Roundtable. Washington, DC. Council on the Teaching of Foreign Languages, Alexandria, VA.

Arce, J. (2000). Developing voices: Transformative education in a first-grade two-way Spanish immersion classroom: A participatory study. *Bilingual Research Journal*, 24(3), 249–260.

Bailey, K., & Freeman, D. (1990). *Creating a community of understanding; creating a community of practice.* Defense Language Institute Foreign Language Center Faculty Development Workshop. Presidio of Monterey, CA.

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

Barsch, J. (2003). *Barsch Learning Style Preference Form*. Texas Center for Adult Literacy. Downloaded from http://www-tcall.tamu.edu/research/NSO/LS/LS_a.htm.

Bloom, B. (Ed.). (1956). *Taxonomy of educational objectives, Handbook I: Cognitive domain*. New York: David McKay.

Campbell, C. (2020). Open architecture curricular design: A fundamental principle of transformative language learning and teaching." In B. L. Leaver, D. Davidson, and C. Campbell (Eds.), *Transformative language learning and teaching* (pp. 43-50). Cambridge, UK: Cambridge University Press.

Campbell, C., & Sarac, B. (2017). The role of technology in language learning in the twenty-first century: Perspectives from academe, government, and the private sector. *Hispania*, 100(*5, Centenary Issue*): 77–84.

Clifford, R. T. (1988). What you test is what you get: Open versus closed instructional systems. *Die Unterrichtspraxis* 21(1): 37–40.

Cohen, B. (2003). *Diagnostic assessment at the superior-Distinguished threshold*. Salinas, CA: MSI Press.

Cohen, B. (2015). *Enhanced final learning objectives activities*. Faculty Presentations at the Defense Language Institute Foreign Language Center. Presidio of Monterey, CA.

Corin, A. (2020). The challenge of the inverted pyramid in attaining Distinguished-level proficiency. *Journal for Distinguished Language Studies* 7: 85-114..

Dababneh, R., & Yuan, R. (2020. accepted for publication). Applications of formative and dynamic assessment in the school of resident education, DLIFLC. In A. Corin, B. L. Leaver, & C. Campbell (Eds.), *Open Architecture Curricular Design: Concepts and courses*. Publisher TBD.

Davidson, D., & Garas, N. (2018). *360 survey*. Unpublished report for the Flagship Culture Initiative.

Davidson, D., & Shaw, J. (2019). A cross-linguistic and cross-skill perspective on L2 development in study abroad. In P. Winke & S. Gass (Eds.), *Foreign language proficiency in higher education* (pp. 217–242). Cham, Switzerland: Springer.

Duri, J. (1992). Content-based instruction: Keeping DLI on the cutting edge. *The Globe* 5: 4-5.

Ehrman, M. (2002). Understanding the learner at the superior-Distinguished threshold. In B. L. Leaver & B. Shekhtman. (Eds.), *Developing professional-level language proficiency* (pp.245-259). Cambridge, UK: Cambridge University Press.

Erhman, M. & Leaver, B. L. (1997). *Sorting out global and analytic functions in second language learning*. American Association of Applied Linguistics Annual Meeting. Orlando, FL.

Ehrman, M., & Leaver, B. L. (2002). *The E&L Cognitive Styles Construct*. Unpublished, copyrighted manuscript.

Ehrman, M., & Leaver, B. L. (2003). Cognitive styles in the service of language learning. *System* 31(3): 393-415.

Farraj, A. (2006). Raising the bar: E-learning for obtaining the Distinguished level in the Arabic language. *Journal for Distinguished Language Studies* 4: 21-30.

Foster, E. (1997). Transformative learning in adult second language learning. *New Directions for Adult and Continuing Education* 74: 33 - 40.

Franke, J. (2020). *Pursuing Distinguished speaking proficiency with adult foreign language learners: A case study*. Unpublished doctoral dissertation. Indianapolis, IN: American College of Education.

Goulah, J. (2006). Transformative second and foreign language learning for the 21st century. *Critical Inquiry in Language Studies: An International Journal* 3(4): 201–221.

Ingold, C. (2002). The LangNet "Reading to the Four" Project: Applied Technology at Higher Levels of Language Learning." In B. L. Leaver & B. Shekhtman (Eds.). *Developing professional-level language proficiency* (pp. 141-155). Cambridge, UK: Cambridge University Press.

Leaver, B. L. (1989). Dismantling classroom walls for increased foreign language proficiency. *Foreign Language Annals* 22(1): 67-74.

Leaver, B. L. (2003a). *Achieving native-like second language proficiency: A catalogue of critical factors*. Salinas, CA: MSI Press.

Leaver, B. L. (2003b). *Individualized study plans for very advanced students of foreign languages*. Salinas, CA: MSI Press.

Leaver, B. L. (2020). Transformative language learning and teaching: The next paradigm shift and its historical context. In B. L. Leaver, D. Davidson, & C. Campbell, C. (Eds.), *Transformative language learning and teaching* (pp. 13-22). Cambridge, UK: Cambridge University Press.

Leaver, B.L. & Atwell, S. (2002). Preliminary qualitative findings from a study of the processes leading to the advanced professional proficiency level (ILR 4). B. L. Leaver & B. S. Shekhtman (Eds.), *Developing professional-level language proficiency* (pp. 260-279). Cambridge, UK: Cambridge University Press.

Leaver, B. L. & Campbell, C. (2015). "Experience with higher levels of proficiency." In T. Brown & J. Bown (Eds.), *To advanced proficiency and beyond: Theory and methods for developing superior second-language ability* (pp. 3–22). Washington, DC: Georgetown University Press.

Leaver, B. L., & Campbell, C. (2020). The shifting paradigm in Russian language programs from communicative language teaching to transformative language learning and teaching. In E. Dengub, I. Dubinina, & J. Merrill (Eds.). *Art of teaching Russian*. Washington, DC: Georgetown University Press.

Leaver, B. L., Campbell, C., Nyikos, M., & Oxford, R. L. (2019). "Transforming the transformers: Breaking through philosophical, cognitive, and emotional barriers. Society, Identity, and Transformation in Language Teacher Education Conference. Minneapolis, MN.

Leaver, B. L., Davidson, D., & Campbell, C. (Eds.). (2020). *Transformative Language Learning and Teaching*. Cambridge, UK: Cambridge University Press.

Leaver, B. L. & Granoien, N. (2000). Философия образования: Почему мы преподаем определенными путями. [Philosophy of education: Why we teach the way we do]. *Философия Образования* [Philosophy of Education] 1(1): 3–9.

Leaver, B. L,. & Shekhtman, B. (Eds.). (2002). *Developing professional-level language proficiency*. Cambridge, UK: Cambridge University Press.

Mezirow, J., & Associates. (Eds.). (2000). *Learning as transformation: Critical perspectives on a theory in progress*. San Francisco: Jossey-Bass.

McClinton, J. (2005). Transformative learning: The English as a second language teacher's experience. *The CATESOL Journal* 17(1): 156–163.

Myers, I. Briggs, with Myers, Peter. (1980). *Gifts differing*. Palo Alto, CA: Consulting Psychologists.

Osterling, J., & Webb, W. (2009). On becoming a bilingual teacher: A transformative process for preservice and novice teachers. *Journal of Transformative Education* 7(4): 267–293.

Oxford, R. (1990). *Language learning strategies: What every teacher should know*. New York: Newbury House Publishers.

Oxford, R. (2016). *Teaching and researching language learning strategies*. London: Routledge.

Poehner, M., & Lantolf, J. (2003). Dynamic assessment of l2 development: Bringing the past into the future. *CALPER Working Papers, No. 1* (October). University Park, PA: Center for Advanced Language Proficiency Education and Research, The Pennsylvania State University.

Poehner, M., & Lantolf, J. (2005). Dynamic assessment in the language classroom. *Language Teaching Research* 9(3): 233-265.

Poehner, M., Davin, K., & Lantolf, J. (2017). Dynamic assessment. In E. Shohamy, I. Or, & S. May (Eds.). *Language testing and assessment. Encyclopedia of Language and Education* (3rd ed.). Cham, Switzerland: Springer.

Robin, R. (2008). Distinguished speakers and screen-time repertoire. *Journal for Distinguished Language Studies* 5: 7-10.

Sandrock, P. (2010). *The keys to assessing language performance: A teacher's manual for measuring student progress*. Alexandria, VA: The American Council on the Teaching of Foreign Languages.

Shekhtman, B. (2003a). Do superior-level students need language instruction? An essay in answer to the myth of natural acquisition and self-study being sufficient at high levels of foreign language acquisition. *ACTR Newsletter* 30(2), 1-3.

Shekhtman. B. (2003b). *How to improve your foreign language IMMEDIATELY*. Salinas, CA: MSI Press.

Shekhtman, B. (2003c). *Working with advanced foreign language students*. Salinas, CA: MSI Press.

Shekhtman, B., Leaver, B. L., with N. Lord, E. Kuznetsova, & E. Ovtcharenko. (2002). Developing professional-level oral proficiency: The Shekhtman method of communicative teaching. In B. L. Leaver & B. Shekhtman (Eds.), *Developing professional-level language proficiency* (pp. 119-140). Cambridge, UK: Cambridge University Press.

Soudakoff, S. (2004). Developing 3+/4 Reading and Translation Skills for Presidential Needs: A Summary. *Journal for Distinguished Language Studies* 2: 19-22.

Spence-Brown, R. (2003). Authentic assessment? The implementation of an "Authentic" Teaching and Assessment Task. Unpublished doctoral dissertation. Melbourne, Australia: University of Melbourne.

The National Standards Collaborative Board. (2019). *World-Readiness standards for learning languages* (4th ed.). Downloaded from www.actfl.org/sites/default/files/publications/standards/World-ReadinessStandardsforLearningLa nguages.pdf.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.