MSI Press LLC
Copyright 2022

Citation:
Corin, Andrew, & Entis, Sergey. (2022). Protocol-based formative assessment: Evolution and revolution at the Defense Language Institute Foreign Language Center. *Journal for Distinguished Language Studies 8*: 95-115.

Protocol-Based Formative Assessment: Evolution and Revolution at the Defense Language Institute Foreign Language Center

Andrew R. Corin and Sergey Entis
Defense Language Institute Foreign Language Center (USA)

Abstract

Protocol-based formative assessment (PBFA) can be a powerful tool for enhancing learning and diagnosing learning challenges. Yet there is an inherent tension between effectiveness and efficiency in the delivery of PBFA. This can be addressed through a variety of strategies: "rationing" PBFA to instances of individual learning difficulties, applying PBFA to all students but in fewer instances, or by engineering greater efficiency into the protocol. Regardless of the strategy adopted, it is taken for granted that PBFA should be maximally integrated with instruction-based formative assessment (IBFA) as an integral component of day-to-day classroom instruction. This article articulates the dilemma as it developed at the Defense Language Institute Foreign Language Center (DLIFLC) between 1989 and 2015 and the path pursued to overcome it through re-design of PBFA.

Keywords: diagnostic assessment; formative assessment; dynamic assessment; zone of proximal development; learner variables; learning styles; text typology; language proficiency; world language education; foreign language learning; Defense Language Institute Foreign Language Center

This article examines the development of formative assessment processes at the Defense Language Institute Foreign Language Center (DLIFLC) in the period between

1989 and 2015 with a focus on *diagnostic assessment* (DA).[20] DA is a protocol-based formative (PBFA) mechanism and an integral component of the concept of diagnostic instruction at the core of DLIFLC instructional strategy. DA has become a valuable tool for the analysis of student learning difficulties, and a fundamental component of efforts to optimize performance of all students. While DA is also borrowed for purposes that may be considered summative, its primary use is as a component of the learning process—to articulate learners' individual characteristics and needs, streamline learning for each, and diagnose and address learning difficulties.

The goals of DA overlap with those of dynamic assessment (as articulated in Poehner, 2008), but are broader in the development and exploitation of "learner profiles" (Sections 2-3, below). In recent years, the attempt to identify and exploit learners' "zone of proximal development" (ZPD, as defined in sociocultural theory and applied also in dynamic testing; see Lantolf & Thorne, 2006; Poehner, 2008) has been incorporated into DA training, thus bringing the approaches closer together.

Section 2 provides an overview of the DLIFLC learning context. It describes features that distinguish DLIFLC from most educational institutions, making it a unique environment for intensive observation of the effects of particular instructional practices. Section 3 sketches the origins and early development of DA. Section 4 examines DA as practiced during the decade through 2015. Section 5 explores how DA protocol was incorporated into instructional practice and quality management during this period, while Section 6 examines DA training. Two paramount lessons emerge from these last two sections. The first, now taken for granted in much of the L2 instructional community, concerns IBFA. This is the understanding that instructors should be constantly attuned to the cognitive styles, strategies, and relevant life experience of each learner, as well as to their current profiles of strengths and weakness viewed dynamically, adjusting activities for both individuals and cohorts to meet current needs. The second lesson concerns inherent tensions between effectiveness and efficiency in PBFA. Simply stated, the time and effort required for effective application of PBFA can become burdensome, while the training and skill maintenance required for effective application can strain or exceed institutional capacity. These tensions can adversely affect PBFA effectiveness as well as instructor and student morale. Sections 7 and 8 outline efforts to overcome this dilemma, and thus achieve the benefits of diagnostic instruction maximally informed by both PBFA and IBFA.

20. This paper was drafted in 2015 and reflects circumstances at that time, during which the authors played leading roles in the development of DA processes and training, culminating in a DA revision project during the years 2014-2015. The insights gained from that project remain valuable today. The authors' leadership roles in DA development concluded following closure of DLIFLC's Diagnostic Assessment Center and reassignment of its tasks to DLIFLC's Language Science & Technology Directorate. They are grateful to Betty Lou Leaver, Steven Koppany, and Bella Cohen for the invaluable recollections they provided concerning the background and development of DA at DLIFLC.

Defense Language Institute Foreign Language Center

During the period in question, some 3,500 students were typically enrolled in DLIFLC's resident initial acquisition (basic) programs in more than two dozen languages, each with 6 or 7 contact hours daily, for between 6 and (more typically) 12 or 18 months. Numerous other students enhanced their proficiencies through DLIFLC's post-basic programs, which included both resident and technology-mediated courses delivered from DLIFLC's home location in Monterey, California. Both post-basic and basic programs were also delivered throughout the world by *language training detachments* stationed at various locations or by *mobile training teams* that would travel to required locations. Other languages with smaller programs were served from DLIFLC's Washington, D.C. office. The graduation objective for DLIFLC basic courses was ILR Level 2+ in listening and reading and Level 2 in speaking, while the graduation requirement was 2/2/1+ in the same skills, respectively. Post-basic courses sought to achieve *global language proficiency* (GLP) at or above ILR Level 3, with an increasing proportion of students exiting at ILR 3+ or 4 in one or more modality.

A key distinguishing feature of DLIFLC's learning environment is accountability for proficiency outcomes. Each student represents a large investment allocated to meet some service or agency requirement. Conversely, students' careers as language professionals depend upon achieving or exceeding GLP standards and other language-related learning objectives. Add to these factors the large number of students processed year-round and the real-world stakes that depend on graduates' abilities, and it becomes clear that DLIFLC must closely monitor the performance of each program and student, taking all possible measures to enhance performance. At course completion students take standardized proficiency tests: Defense Language Proficiency Test (DLPT) in reading and writing, and the Oral Proficiency Interview (OPI).

Yet high stakes and high-stakes exams do not ensure success. For this reason, formative assessment played an ever-increasing role in the DLIFLC learning process during this period, becoming integrated, in differing ways, into both basic and post-basic DLIFLC courses.

Development of Diagnostic Assessment (DA)

DA arose at DLIFLC out of the practice of providing formative feedback based on immediate recall protocol (in the sense of Bernhardt & James, 1987) and learning styles beginning around 1989. The term "diagnostic assessment" was applied later, in 1997-1998, when it was assessed that what was missing from formative assessment processes at DLIFLC was utilization of the ILR scale for formative purposes through a standards-based matrix, and foundations were laid for a DA protocol encompassing the modalities of listening, reading, and speaking.

The new protocol included a diagnostic interview by two native speaker interviewers, who would present authentic content and elicit a ratable and analyzable sample, somewhat in the manner of an oral proficiency interview. Selection of materials was governed by text typology (TT) in the sense arising out of Child (1987), which roughly means establishing

the ILR level of any reading text based on criteria analogous to those used to establish the ILR level of a ratable speech sample. TT is a discrete skill requiring norming. DA interviewers would use a "package" of texts, each rated at a specific ILR level and each associated with a unique set of comprehension questions requiring proficiency at the specified level. Interviewers would adapt the interview up or down depending on the proficiencies demonstrated by the learner. Initially, assessment teams were formed for Russian, Spanish, Korean, Chinese, and Arabic, with Persian Farsi added in 2000. Two DA specialists were allotted per language. The developers were aware of the range of extra-linguistic individual variables (cognitive styles, etc.) relevant to L2 learning, but did not feel that it was feasible to encompass them within the initial DA protocol.

The goal of the DA interview was not merely to determine GLP levels and specific strengths and weakness; analogous to dynamic assessment, it sought to identify areas that learners needed to address in order to achieve the next level of GLP, focusing on those within their immediate grasp (not yet referred to at this point as ZDP).

Rating and analysis of elicited samples and development of individualized learning plans (LPs) was based on the ILR Level Descriptions for each skill.

Following its inception in 1998, DA was offered on an on-demand basis to DLIFLC schools and external clients; it was recommended to basic-program schools that they perform one DA for each student at the end of the second semester (of a three-semester program), in time to address lacunae and make adjustments. Some department chairs preferred to perform DA earlier, at the end of the first semester.

The fact that DA was based on the ILR scale (like the DLPT) stimulated popularity, and numerous clients came forth, some using DA for purposes (placement, syllabus design, materials selection, program validation, etc.) beyond its intended purpose of enhancing efficiency of the learning process through individualization of instruction.

The popularity of DA, however, soon led to resource issues that limited DLIFLC's ability to apply DA as broadly and effectively as planners envisioned. One major issue was the labor-intensiveness of the process. When performed according to specifications, each DA could require as much as 13 hours of effort. The initial two-person teams per language could not meet demand, and routine use of multiple iterations of DA within a course clearly could not yet be anticipated.

The labor-intensiveness of DA gave rise in 2002 to idea of online diagnostic assessment (ODA), since it was felt that face-to-face DA couldn't meet demand on a cost-effective basis. ODA for reading and listening is now available through the DLIFLC internet site for many languages, though face-to-face DA remains the preferred model when feasible, allowing for more articulated and individualized analysis and treatment.

Yet another deleterious effect arose out of efforts to cope with the labor-intensiveness of the process. In response to tedium and time pressure, a tendency arose for DA specialists to take short cuts, e.g., "templatizing" individual LPs, so that the crucially individual aspect of the LP was lost, devolving into generic statements. This difficulty was exacerbated by the limited ability of some DA specialists, native speakers of the target

language, to compose feedback and recommendations in adequate English, especially as learning approached higher levels.

From the point of view of the teaching departments, allotting time for DA from an already tight program was inconvenient and could be evaluated as counter-productive if evidence of a positive effect on learning outcomes was not manifest. This became likely if DA was performed in a rushed, pro forma manner, leading to a vicious circle. The desire of departments to work DA into their programs in the most efficient manner led in some cases to an especially pernicious effect which undermined DA's entire premise: since evaluation of a learner's current GLP level was one aspect of DA, some departments began to use DA for summative purposes, e.g., in lieu of unit or end-of-semester tests. When this occurred, DA came to be perceived by students as a summative measure and was received by them with all the enthusiasm usually adhering to tests.

In 2000 the School of Continuing Education (SCE) was created to house DLIFLC's intermediate and advanced courses, distance learning, and a range of outreach programs, including DA. DA development and delivery proceeded for some two years before going dormant for lack of funding. One major advance during this period was the regular provision of DA in support of the Defense Threat Reduction Agency, whose interpreters are trained by DLIFLC.

During fiscal year 2006, program budget decision PBD 753 provided funding for eight DA specialists for the Directorate of Undergraduate Education (UGE, which encompassed DLIFLC's basic-program schools) and eight DA specialists for what was now the Directorate of Continuing Education (CE), leading to a revival of DA. In June 2006, the eight DA specialists assigned to CE were combined into a Diagnostic Assessment Center (DAC). DAC was mandated to follow a new model of DA development and training, combining existing DA protocol with the earlier concept of formative assessment based on individual learner variables. The resulting enhanced DA protocol generated a *learner profile* based on multiple parameters of individual learner variation, together with a *linguistic profile* based on a three-skill interview, yielding an LP following the model developed in 1998 forward.

DA training also followed a new model, using the DAC as a force multiplier to spread DA skills throughout DLIFLC. Instead of focusing primarily on development and delivery of DA services to CE, the DAC would develop both DA protocol and non-language-specific models of DA training, devoting most of its energy to training and certifying instructors from all languages and all DLIFLC teaching schools as DA specialists. The goal was "to turn 8 into 800," freeing schools from dependence on small teams of DA specialists and ensuring that DA would be conducted by the same instructors who taught the students.

The Enhanced Diagnostic Assessment Protocol

DA elicits and interprets data on multiple aspects of a learner's individual characteristics, relevant biographical detail, and proficiencies. Products include a profile of learner characteristics, profile of linguistic proficiencies and needs, and individualized LP. Together these provide an informed assessment of learners' current needs, to be

updated as required.

DA is comprised of three stages: pre-interview data collection; three-skill interview; and post-interview follow-up, during which profiles are constructed and, based on them and in consultation with the learner, a tailored LP. At course completion, the learner becomes the author of the LP.

Pre-Interview Data Collection

Three forms of data are collected, bringing together social, psychological, and linguistic information. This portion of DA need not be repeated in subsequent iterations. First, a biographical questionnaire elicits information on aspects of learners' current situations and life history found historically to affect learning outcomes (e.g., Ehrman, 1996b, 164-172). Second, information is compiled on parameters of individuality relevant to the effectiveness of particular manners or modes of learning.

Personality type, as it affects learning contexts, modes, and tasks in which a learner will flourish, is assessed based on Jungian models (perhaps the best known of which are the Myers Briggs Type Indicator and Keirsey Temperament Sorter). DA training in this area is based on Leaver et al. (2005, Ch. 4, especially pp. 113ff.) and Ehrman (1996a).

Cognitive styles are assessed using the Ehrman and Leaver Learning Style Questionnaire Version 2.0 (E&L; cf. Ehrman, 1996b; Ehrman & Leaver, 2002, Leaver et al., 2005). E&L defines a global dichotomy of cognitive styles: *synopsis* vs. *ectasis*. *Ectenics* are learners who strive for conscious control over perception, information processing, and learning (i.e., a deliberative approach to learning), while *synoptics* are comfortable relying on less conscious or less controlled processing, willing to "trust their gut" (osmotic approach to learning; Ehrmann & Leaver, 2003, p. 395). The synopsis vs. ectasis dichotomy is articulated into 10 subordinate categories (cf. Table 1).

Synoptic Learning	Definition	Ectenic Learning	Definition
field independent	selects from context	field dependent	relies on context
field sensitive	learns by osmosis	field insensitive	learns by discrete item analysis
random	follows an internally developed order of processing	sequential	looks for an externally provided order of processing
global	focuses on the big picture	particular	focuses on the details
inductive	goes from examples to rule	deductive	goes from rule to examples

synthetic	prefers to assemble information into new wholes	analytic	disassembles wholes into constituent parts
analogue	metaphoric thinking	digital	linear thinking
concrete	interacts with the world directly and learns through application	abstract	interacts with the world through cognitive constructs and learns from formal rendition of knowledge
leveling	looks for similarities	sharpening	looks for disparities
impulsive	responds first	reflective	thinks first

Table 1. E&L Learning Styles Sub-Categories. Cited from Corin & Leaver (forthcoming).

Tolerance of ambiguity (in the sense described by Ehrman, 1999) is another parameter with profound implications for the forms of course organization in which learners may flourish or, conversely, fail regardless of capability and effort. This is discussed during DA training, but to date no feasible manner has been devised to encompass it systematically in DA protocol. Instead, tolerance for ambiguity is deduced or predicted from cognitive styles, relying on the untested assumption that they are correlated, or is elucidated through consultation with the student.

Sensory preferences. As generally acknowledged, learners differ in the ease or comfort with which they utilize various sensory channels (visual, auditory, tactile, kinesthetic) for processing information. A learner with strong visual preference and low auditory preference, for example, may be more comfortable learning from written (i.e., visualized) texts than from auditory (heard) texts, etc. Learners' sensory preferences are assessed using the Barsch Learning Style Inventory (Leaver & Oxford, 2001-2005, p. 87). An overview of this topic as it relates to second language acquisition, albeit focused on the VARK questionnaire, is Fleming and Mills (1992), which is used in DA specialist training.

Motivations. Alongside anecdotal information obtained through the biographical questionnaire, these are assessed using the Motivated Strategies for Learning Questionnaire (cf. Pintrich et al., 1991; Stoffa et al., 2011 on interpreting the MSLQ; Dörnyei & Csizér, 1998 has been used in DA training).

The results of this data collection are compiled in chart form, providing an easily interpretable approximate learner profile, which requires confirmation and adjustment through observation and consultation with the learner.

The third component of pre-interview data is an L2 writing sample composed without the use of linguistic reference materials, typically on a topic related to the learner's life history or future plans. Elicitation varies depending on a preliminary estimate of the learner's GLP. In addition to linguistic data instantiating areas of strength and weakness, the sample may provide information on motivations or learner variables beyond those

elicited through other instruments. The writing sample is not a ratable sample. Rather, it is used much like the warm-up segment of an OPI—to provide assessors with a sample exemplifying the learner's current proficiency level and some specific strengths and weaknesses. It suggests an appropriate opening level for the DA interview and some areas to be probed during elicitation.

The Three-Skill DA Interview

The DA interview, as currently structured, consists of three components: listening, reading, speaking. Its goal is two-fold. First, it attempts to identify the learner's current GLP on an incremental scale based on the ILR, identifying a "floor" level of sustained performance and "ceiling" level (floor + .5) at which performance cannot be fully sustained. Second, it identifies and explores areas of weakness to be articulated and addressed through subsequent intervention. It focuses on learners' proximal learning potential, identifying tasks they can carry out with modest facilitation, in distinction to those they can carry out without facilitation and those they cannot carry out without extensive support.

The speaking interview is modeled on the structure of the ILR OPI but follows a distinct protocol reflecting the DA's formative (sympathetic) goal in contrast to the OPI's neutral summative assessment. It is comprised of the same inventory of tasks utilized in the OPI, ranging from *short questions* and *simple short conversation* to Level-4 tasks requiring intricate preludes: *abstract topic*, *support opinion* and *hypothesize*. However, where the OPI requires a strict protocol of alternating *level checks* (to establish/confirm the *floor*) and *probes* (to establish the *ceiling*), the formative purpose of the DA interview requires a greater emphasis on probes. In later portions of the interview probes may follow one another in succession, in order to explore as many areas of weakness as is feasible. Tasks assess learners' proficiency in terms of the same ILR performance domains as the OPI: global tasks and functions, text types produced, lexical control, structural control, delivery, sociolinguistic competence. Although the interview is not intended primarily to establish a reliable ILR level, the sample is rated, and experience with near-end-of-course DAs demonstrates that well-normed DA specialists typically assign ratings in close accord with OPI ratings assigned soon afterward.

For the reading interview, as indicated above, the DA specialist uses a packet of texts encompassing the full range of ILR levels, each text at a defined level verified by TT. The texts span a variety of topical domains, which may allow the specialist to identify areas of domain fossilization (in the sense of Ehrman, 2002 and CDLC, 2008, pp. 41-44) inhibiting achievement of upper ranges of proficiency. The comprehension questions associated with each text should be understood more broadly than the usual classroom sense; at higher levels they may probe comprehension of such attributes as non-explicit expression of authorial intent and authorial "voice."

The interview itself typically includes three or four texts, depending upon the course of the interview and the learner's fatigue level. It is adaptive, the level being adjusted up or

down as required. So as to focus the assessment on reading (rather than oral) proficiency, the interview is conducted in English.

The listening interview follows the same pattern as the reading interview, based on a packet of TT-rated listening passages, each with its set of comprehension questions. Each text is played twice, and the learner may take notes.

Post-Interview Follow-Up

Post-interview feedback is based on two primary elements:

- *Learner profile*: pre-interview questionnaire results collated in chart form. This is accompanied by an overview based on quantitative and qualitative data.
- A *linguistic profile* of the learner's current proficiencies, indicating GLP but focusing on specific areas of weakness within the learner's ZPD. The GLP rating is one aspect of learners' proximal learning potential, and it is also a useful indicator of recent learning when DA is applied on successive occasions.

Based on the profiles, a tailored LP is compiled, then shared and explored with the learner during a post-interview consultation before finalization. A primary goal of the consultation is thus to fine-tune the LP, to help learners focus efforts to achieve their proximal learning potential. LPs should not be extensive but must go beyond broad generalizations to include actionable recommendations and may include benchmarks toward success and possibly a timeline. They may contain recommendations on two tracks—accommodation and confrontation, assisting learners to capitalize on their favored strategies, sensory channels, etc., while making them aware of alternatives and helping them to expand their repertoire of learning strategies. LPs may initially be composed by the DA specialist in collaboration with the learner, but by course conclusion they should be composed in L2 by the learner.

Beyond assisting learners to understand their proximal learning needs (i.e., for the immediate upcoming period), a second goal of the post-interview consultation is to discuss the learner profile (cognitive styles, sensory preferences, etc.) based on the evidence of both the questionnaires and three-skill interview. Questionnaires utilized in pre-interview data collection are far from infallible, so it is important to identify areas in which results are not in accord with observation or learners' beliefs.

All results are shared fully with all members of the instructional team and become the cornerstone on which tailoring of instruction will be based.

Finally, a class profile should be composed, revealing the predominant characteristics of the learner cohort and allowing instructors to tailor instruction based on predominant types with individualization to meet the needs of all. For example, a class with a high proportion of introvert ectenic learners may require more gradual work-up to stimulate performance in scenario-based activities. A class in which extrovert kinesthetic learners predominate may require adaptation of a program designed originally around content-based area studies emphasizing lengthy readings, individual analysis, and production in formalized genres such as diplomatic "requests for information."

Utilization of Diagnostic Assessment

Within DLIFLC's eight resident basic-program schools a variety of practices prevailed during the period through 2015, with a gradual trend away from individual as-needed application of DA toward systematic application. Within CE's School of Resident Education (RE),[21] DA was most thoroughly integrated into the learning process during the period 2006-2015, with systematic delivery of multiple iterations of DA in all courses. Remarks in this section focus on the RE's intermediate/advanced programs, with offerings in 8 languages as of 2015.

RE courses did not utilize textbooks. Curriculum was laid out in program/course descriptions and implemented primarily through content-based and scenario-based instruction using authentic materials (materials produced by, and for use by, native L2 speakers). Int/Adv courses ranged in length from 4-12 months,[22] with relatively few traditional quizzes and tests. DA thus bore much of the burden for in-course formative assessment. Except in the shortest (12-16 week) courses, students would have three DA iterations: at course outset, mid-point, and one month from completion.

Initial DA was particularly important in post-basic instruction. Students' linguistic profiles had diverged already during initial acquisition; subsequently they went on to a variety of other activities prior to their enrollment in CE courses, leading to further divergence of both the range and level of their proficiencies, resulting in a wide variety of class profiles.

Mid-point DA assessed progress toward learning objectives, set new baselines for GLP, specific knowledge, and proficiencies; identified major learning difficulties for intervention; and assisted in program correction.

Near-completion DA flagged any glaring lacunae requiring last-minute intervention, and students, in consultation with instructors, prepared their individualized LP for post-course learning. Since RE courses were conducted in L2 immersion mode throughout, the LP was composed in L2, reinforcing learners' confidence that they could utilize L2 to meet real-world needs. An LP compiled by the learner was also considered more likely to be implemented. Near-completion DA ratings that diverged significantly from out-going DLPT and OPI results were analyzed to determine what interventions or remedial actions might be required.

In addition to the three required iterations, all students (as indeed all DLIFLC students) who participate in overseas immersion programs underwent DA immediately before and after immersion. This assists in evaluation of an immersion's immediate benefit (bearing in mind that far from all learning benefits are measurable immediately upon return) and sets new baselines for learning following the "boost" the immersion hopefully provided. This is particularly important because Int/Adv immersion programs were content-based, conducted in a university (not a language-school) setting, and designed to simulate the

21. Since this article was drafted, the resident intermediate-advanced-course component of CE has been transferred to UGE.

22. Shorter courses (12-16 weeks) introduced in 2014 limited DA delivery to two instances per course.

experience of native-speaker university students studying the same topics. Such 2-4 week immersion programs commonly propelled well-prepared students over a GLP cusp with gains of .5 level for listening, reading, and speaking in as little as two weeks.

RE Instructors incorporated DA results into instruction in multiple ways. First, the breadth of learner characteristics and linguistic profiles was reflected in lesson planning through adaptation toward the cohort while ensuring individualization, breadth of activity types, variation of pace, etc., to meet all learners' needs. Second, DA results were a component of the detailed records that tracked each student's and class's progress. They were discussed at team meetings, used to plan intervention, were a crucial component of every course outcome analysis, and informed monthly individual student counseling. Finally, DA was a keystone of diagnosis and treatment of learning difficulties—the first item reviewed when CE academic specialists were called upon to assist in cases of particularly intractable difficulties.

Diagnostic Assessment Training

DA training has three primary goals. First, it creates and maintains a corps of DA specialists proficient in applying DA protocol. The second, broader, and arguably more important in the long run, is to enable DA trainees, in their day-to-day role as instructors, to better understand their students and more effectively facilitate their learning—developing their realization that classes are composed of learners with vastly differing profiles and needs and sensitizing them to strategies to more effectively address individual need. The third is to promote DA's purpose of helping instructors reduce attrition and profound learning frustration through a partnership of learners and learning facilitators (instructors and other support personnel).

DA trainees must master a range of conceptual components, including:

- dimensions of learner variation, how they interact to produce infinite composite profiles, and implications for the process and outcome of learning,
- the concept of proximal learning potential,
- the differential relevance of DA for initial acquisition and upper-range learning, and
- varieties of fossilization inhibiting achievement of near-native proficiency.

The last two areas reflect recent developments in training not yet fully implemented as of 2015. Training goes on to include skill sets beyond assessment of learner variables, including TT and elicitation, both of which require norming. DA practitioners must also understand intervention through both accommodation and confrontation, including composition of tailored LPs. Based on experience at both DLIFLC and the Foreign Service Institute (Betty Lou Leaver, personal communication), it has been estimated that about one year of practice is required for instructors to develop proficiency in the recognition and assessment of learner variables, elicitation, rating, and intervention necessary for effective formative assessment.

This breadth of topics and depth of training and norming has made it difficult to meet training objectives. This includes training enough DA specialists, training them to

the point of proficiency, and keeping them sufficiently normed to provide reliable ratings, assessments of learner variables, diagnoses, and interventions. As a result, it has proven challenging to integrate DA as an integral component of all DLIFLC courses (especially in the much larger basic programs), much less to approach the goal of universal DA training for all instructors.

To address these challenges, the DAC experimented with various training models, initially 2-4 week intensive certification workshops, later blended media models combining asynchronous technology-mediated training with two three-day face-to-face norming workshops. Annual language-specific refresher and norming workshops for previously trained specialists were provided. The DAC simultaneously approached the broader goal of a DA-informed instructor corps through DA familiarization training for a larger number of instructors, initially through one-week workshops, later adding an on-line course.

Ultimately, the DAC adopted and began to implement a train-the-trainer (TtT) approach, distinguishing four levels of training. At the highest level, DAC staff required about a year of on-the-job performance to develop proficiency in training trainers. The second level was that provided by the DAC staff to prepare school-based DA trainers who would deliver DA specialist training (third level) to instructors in their respective schools. A fourth level—DA familiarization—was offered online by the DAC.

The DAC ultimately provided DA certification training to 509 instructors in 22 languages, certified 215 DA specialists (19 languages), and provided familiarization training to another 184 instructors (14 languages). Though a significant success, out of a faculty of almost 1,500 it was far from adequate when one considers faculty attrition, movement, and the ever-present requirement of refresher training. TtT should, in principle, allow a significant increase in this number, but the reality is more limited. On the one hand, TtT brings the danger of watering down training. On the other, aside from the issue of training resources there is also that of instructor resources. In the intense DLIFLC teaching environment, hours when instructors are not available for teaching or course/class preparation are difficult to budget, and DA training must compete for this time with yet other responsibilities, including numerous annual training requirements.

The overarching conclusion concerning DLIFLC's experience with DA through 2015 is that the process is highly effective when carried out systematically in all of its aspects. While its effects cannot be isolated from other factors, it is clearly one of several measures that resulted in a more than 50% increase in the percentage of Int/Adv students achieving proficiency requirements in all skills over the period 2006-2013. Despite these gains, the goal to universal application of DA and its benefits remains elusive.

Addressing the Effectiveness vs. Efficiency Dilemma

As articulated above, DA places heavy resource demands on all involved parties. Where insufficient training or short-cutting due to time pressure leads to inaccurate or generalized results or a failure to follow up on DA, the absence of positive results undermines confidence in DA on the part of students, instructors, and administrators

alike. This creates a vicious circle, with schools or instructors reluctant to devote resources to DA and DA training. Insufficient awareness can also lead some instructors and managers to perceive DA as a last-gasp intervention technique to save students at risk, rather than a holistic approach to facilitating learning from day 1.

There are at least two related underlying tensions that contribute to this conundrum undermining DA's potential—one related to the instrument itself and the other to faculty training. First, the original DA design utilized "off-the-shelf" components that had been designed for testing purposes—OPI for speaking proficiency, and TT associated with level-rated content questions for reading and listening. This created a dichotomy between learning and formative assessment processes, through the use of instruments and procedures beyond those employed in the learning process *per se*. This tension can be addressed in part through redesign of DA protocol, which will be discussed below.

The second underlying tension derives from the first. Because DA involves instruments and procedures outside the learning process *per se*, additional training is required beyond that required for instructors to become proficient in facilitating learning (i.e., teaching). This additional training encompasses numerous concepts and techniques of practical application, more than can be covered comfortably in a limited workshop. It is more like a university program squeezed into a single workshop. Developing proficiency in DA implementation also requires extensive time-on-task.

The ultimate roots of these tensions are thus clear. On the one hand, DA training, involving as it does both formative assessment *per se* and exploitation/intervention based upon this assessment, should arguably be a central component of any L2 instructor training program. Viewed in this light, the fact that DA development and training as of 2015 were separate from the Instructor Certification Course taken by all new DLIFLC instructors would be one of the greatest impediments to the effective and efficient implementation of DA at DLIFLC. On the other hand, the fact that DA, though intended to inform the learning process, is carried out external to that process based in part on components designed for testing purposes, creates an inherent inefficiency in the use of DA. An effective strategy for resolving both of these tensions would thus involve redesigning DA to fully integrate protocol-based formative assessment (PBFA) as a component of the learning process *per se*, so that time devoted to PBFA is simultaneously time devoted to learning rather than subtracted from it. Only in this way can training for PBFA be integrated with IBFA into a holistic approach to instructor training uniting three essential components: 1) understanding the linguistic and extra-linguistic realities to be mastered, 2) understanding each student's individuality and the dynamics of learning cohorts, and 3) mastery of strategies and techniques to optimally facilitate learning by individual learners within their cohort.

Ideally, instructors should be able to evaluate each learner's unique learner and linguistic profile and respond to individual and cohort needs based on observation alone, without the need for PBFA. This level of proficiency is unlikely to be achieved by most instructors, so PBFA will continue to provide invaluable scaffolding, allowing instructors to gain a sufficient appreciation of their learners' needs and structure effective responses

to those needs. DLIFLC therefore, as of 2015, followed a dual track approach to PBFA. The first was by a TtT approach to allow maximal access to existing DA. The second was a simultaneous effort to redesign DA itself with the goal of re-integrating DA into the learning process.

Redesigning Diagnostic Assessment

The goal of redesign was to move DA away from components rooted in testing modalities toward a structure designed for formative assessment as a component of the diagnostic instructional process. Instructors regularly practice *ad hoc* IBFA through every-day observations, optimizing support for learners through individualization while adapting group activities based on the profile and dynamic of the cohort. To maximize its effectiveness as well as its efficiency, DA must be maximally integrated with these day-to-day classroom IBFA processes.

The focus of the redesign was thus on the three-skill interview, the core of the linguistic assessment. The pre-interview data collection, in contrast, does not require class time; questionnaire results are easily quantifiable in chart form; and the writing sample is easily incorporated into the learning process. Post-interview processing is thus amenable to streamlining and, if DA can be carried out by any instructor, can be largely integrated into the instructional process.

As envisaged, re-design of the DA interview included four conceptual innovations:

1. expansion of text genres used in assessment,
2. utilization of multi-level texts,
3. RC and LC elicitation based on performance tasks rather than comprehension questions (based on the axiom that "a well-designed task cannot be completed without text comprehension;" "Thoughts for Thursday," 5 June 2014, DLIFLC Provost's Office), and
4. rating based on *proficiency cusp tables*, which require less training for effective use than ILR Level Descriptions.

Genre Expansion for Reading and Listening

As of 2015, internet news media were the main source of DA text selection, in part because DA was then limited to specific area-studies domains considered central for DLIFLC students, and relevant texts are readily available from these sources. Another reason is that one can readily find texts in these sources that are easily TT ratable (i.e., having a single unambiguous and consistent ILR level). In real life, however, texts typically do not fall neatly into discrete levels, and GLP in reading and listening requires the ability to process a broader range of genres and topics. Adherence to topical restrictions and TT guidelines provides a clear and transparent structure for DA but reduces the usefulness of materials for assessing GLP while increasing the difficulty of creating materials. The problem is thus that these selection criteria, useful in a testing environment, detract from the usefulness of DA for formative assessment.

Removing limitations on genres and topics:

- allows for assessment of a range of L/R proficiencies more truly representative of conditions encountered in real life (i.e., of true GLP),
- assists in the identification of domain fossilizations (in the sense of Ehrman, 2002; CDLC, 2008), and
- integrates DA with the learning process, since DA packets (texts and tasks) are selected using the same criteria as used in selecting authentic materials and tasks for classroom GLP development.

For assessment at higher levels, expansion of genres opens up an Eldorado of linguistic features available for analysis and interpretation, from simple grammatical structures to metaphor, nuance, hidden agendas, discourse virtuosity, diversity of voices and accents, etc. Expansion of genres also promotes thinking out of the box on the part of instructors/DA specialists as opposed to following a script.

Genre expansion may encompass short stories, letters, diaries, social networks, poetic recitals, plays, excerpts from movies and talk shows, etc. To be sure, this broader range of text types requires more highly honed elicitation skills, but these are essentially the same skills as required for effective facilitation of classroom-based learning utilizing the same materials.

Multi-level passages vs multiple single-level passages

It has been observed over the course of many DA interviews that results can be skewed due to learner disorientation arising from discontinuity of content and assessment topics. This occurs when the assessor moves from one text to the next, with a different topic, level, and set of questions. This transition entails a discontinuity of cognitive processes, which can limit learners' performance and affect the assessor's interpretation of that performance. The skewed interpretation, in turn, can lead to the proposal of ineffective learning strategies.

Re-designed DA therefore used *multi-level texts* for assessing RC and LC, each united by a cohesive idea, but supporting tasks that require a variety of proficiency levels. The DA interview is adaptive and dynamic in nature, adjusting based on the flow of the interview. Elicitation based on multi-level texts allows assessors to adapt questions or tasks upward or downward without disorienting shifts in the learner's focus. This also allows for more natural deviations from the pre-set questions or tasks and should yield a more revealing interaction and accurate impressions. Multi-level texts are more representative of those typically encountered in real life, thus allowing for a more realistic assessment of GLP. Finally, use of multi-level texts makes it possible to decrease or eliminate TT training and periodic refresher norming.

Comprehension questions vs reading/listening performance tasks

The oral proficiency interview component of current DA is comprised of a regime of performance tasks, while the RC and LC interviews (like the DLPT) require learners to

provide the most appropriate response to multiple-choice (MC) comprehension questions. Two tools with different potentials and limitations are thus applied to elicit samples from which the assessor will extract a unique interpretation. It is not known whether this disunity of method skews results, but at best it injects a further element of uncertainty into the assessment process.

MC questions, while arguably an effective tool for eliciting responses requiring particular levels of linguistic proficiency, are hardly analogous to typical real-life interactions with R/L texts and are therefore inappropriate for PBFA if it is to be maximally integrated with IBFA. Recent literature, for example Wiggleworth's (2008) survey, albeit focused more on summative assessment, lends general yet critical support to the claim that on the road to GLP, "reading tasks and assessments should mirror real-world uses of reading" (Davis, n.d.) and, of course, listening. Re-designed DA thus assessed R/L through performance tasks, with the caveat that in assessing GLP one cannot fully observe a distinction (noted by Wigglesworth, 2008) between focus on successful task performance vs. focus on the elicited linguistic sample. Task-based assessment is expected to provide a truer appreciation of learners' proficiencies, both in level and particulars, while overcoming any deleterious effects of the current methodological inconsistency. Tasks can include generating or answering questions; non-linguistic tasks; filling out charts, grids, or semantic maps; indicating structural aspects of the text; providing meta-cognitive judgments; and product-oriented tasks. In other words, they can span the range of tasks that an instructor might otherwise use for learning purposes. Mastery of required elicitation skills should therefore require only modest training beyond that required for effective task-based classroom instruction (e.g., in gradated facilitation to achieve the dynamic aspect of assessment), and largely bridge the gap between interview and the classroom.

A third advantage of the task-based approach is that, unlike comprehension questions, performance tasks allow for observation of the learner's cognitive styles, strategies, etc. in action. This supplements data elicited by questionnaire and consultation, thus assisting the assessor in providing a more useful individualized LP.

For rating purposes, one or more tasks will be provided at each ILR level supported by a given text. To maximize their effectiveness for rating purposes, they will be applied in an alternating pattern of level checks and probes, as in the OPI elicitation regime.

Streamlining DA Reference Materials

Within the pre-existing DA protocol, assessment of GLP level requires DA specialists to apply a set of reference charts entitled "Description of Performance by ILR Level" for the various skills, performance domains, and levels encompassed by the ILR. These are applied to the sample elicited during the DA interviews to compile some 12-14 pages of calibrated ratings and feedback. This requires detailed knowledge of the descriptions for various skills, performance domains and levels, as well as intensive norming and periodic refresher training both for elicitation skills and rating. When carried out comprehensively, the process can be sufficiently lengthy that results no longer seem fresh when shared with

the student and teaching team; when this occurs, enthusiasm to apply the fruits of the effort may have dissipated.

Re-design of DA required an assessment reference tool that is cohesive and user-friendly, yet retains or increases depth of substance, reflecting multiple parameters of linguistic performance. It was decided that the most promising approach—simplest yet most comprehensive and nuanced—was through what have become known as *proficiency cusp tables*. These are created by translating the criteria from the various ILR performance domains (lexical, structural, socio-linguistic competence, etc.) into language-specific tables of "can do" statements for particular abilities at particular ILR levels. These in turn are based on an unpublished non-language-specific draft developed under the auspices of the National Foreign Language Center and entitled "Proficiency Cusp Tables or the Generic Learning Profiles (GLPs) for Reading and Listening" (Betty Lou Leaver, personal communication; see now Leaver, this volume). This instrument dissects linguistic performance into 10 multi-dimensional categories not limited strictly to ILR. Categories relevant to formative assessment of GLP beyond ILR include "Strategic and Emotional Competence" and "Attentional Focus." The cusp tables are associated with answers and keys for determining the profile of a learner in a granulated manner. Through their categories, they prompt DA specialists/instructors concerning perspectives from which enhancement of a learner's performance can be approached. This approach was in fact adopted by earlier DA developers, and cusp tables developed for several languages. However, the approach was dropped, presumably because resources were not available for development of cusp tables at all proficiency levels for all DLIFLC languages.

Cusp tables have deficiencies which must be addressed before they can be generally applied as the basis for assessing GLP. Several LC categories are incomplete, and there are at present no speaking and writing tables. Until the latter are developed, ILR level descriptions must remain in use. Cusp tables might first be applied for LC/RC, the difference in criteria corresponding to the dichotomy productive vs. receptive skills.

8.5. Assessing Components of Redesigned DA

The new approach to DA underwent its first trial applications for Arabic, Chinese, English, Korean, Russian, and Spanish, using post-basic students (ILR 2-3) as subjects. The trial assessments were carried out within the context of a DA Summit held in February 2015. The trial DAs employed multi-level texts, expanded genres, and task-based elicitation. Cusp tables were not employed, rating and analysis being based on ILR level descriptions. Overall, the three components that were tested were positively evaluated.

Use of multi-level texts was positively assessed by both learners and assessors, as this brought the assessment process closer to the classroom learning process, while allowing for the identification of learners' strengths and weaknesses. The assessors also appreciated the fact that multi-level texts eliminate the need for cumbersome TT. Nevertheless, use of multi-level texts for assessment raised one unanticipated challenge. Multi-level texts tended to be longer, which in turn increased the role of memory vis-à-vis proficiency as factors limiting performance, necessitating compensatory strategies in formulating tasks and questions (e.g., drawing attention to particular portions of a text). However, it was also

noted that any increased "cognitive load" of multi-level texts also increases opportunities to observe cognitive styles, which is desirable for formative assessment.

Use of expanded genres was positively evaluated by learners, who perceived the experience as a refreshing expansion of their learning into new areas.

Task-based elicitation was assessed as being useful for identifying specific strengths and weaknesses. Learners' cognitive styles and strategies, moreover, were observable during the task-based assessment process, as was anticipated, so that the interviews complemented learner-profile data elicited by questionnaire. A further advantage that had not been anticipated was that weaknesses in socio-cultural proficiency emerged far more clearly in some instances than would have been the case using comprehension questions.

Task-based assessment of GLP level proved a greater challenge. It was felt that in future trials the levels of the tasks provided for each text should be made more transparent. The use of cusp tables is also expected to mitigate this difficulty, as they provide exemplification of performance characteristic at particular proficiency levels. By the same token, cusp tables should also make it easier to design tasks.

Though use of English for elicitation was retained in the task-based R/L assessment—considered a necessary evil for rating purposes—interview teams believed that this artifact of the testing environment detracted from the goal of integrating PBFA into the learning process. This was especially true for the post-basic student subjects in the trial DAs. Accurate assessment of reading and listening proficiencies without use of English requires techniques to isolate learners' proficiencies in listening and reading from those in speaking, so that deficiencies in the latter do not mask proficiencies in the former.

In some trial interviews the dynamic aspect of assessment was achieved: it became clear whether the learner was "close to," or "far from," the ability to independently complete a task. To ensure that this occurs and provide a dynamic aspect of quantification or gradation to the rating process, a schedule of graded levels of scaffolding (support or hints) might be prepared for each task, roughly along lines suggested by Aljaafreh and Lantolf's (1994) "Regulatory Scale" of mediation.

The re-designed DA with task-based elicitation, multi-level texts, and expanded range of genres was positively evaluated by all student subjects, all of whom had previous experience with the pre-existing DA. They were animated by the process, in part because the tasks engaged higher-order learning skills, and in part because they were actually learning. Beyond the knowledge gained from manipulating textual content, they were becoming aware of specific limitations in their proficiency and overcoming some on the spot, so that they ended interviews with greater knowledge or proficiency than they possessed at the start. Interview teams for the various languages also agreed with this assessment.

It was agreed that future trials of re-designed DA should utilize cusp tables, at least for those languages and skills for which they are available, as it was anticipated that this would significantly facilitate analysis and GLP rating.

In conclusion, the three "cornerstones" of re-designed DA that were tested—genre expansion, multi-level texts, and task-based elicitation—received a generally favorable

evaluation, with suggestions for improvement. The greatest remaining challenge was in the measurement of performance, the answer to which lies in the fourth cornerstone—language-specific proficiency cusp tables as reference tools.

Conclusions

DA is a highly effective PBFA tool for enhancing learning outcomes when fully and proficiently applied. However, as the scale of its employment increases, DA's usefulness tends to decrease due to issues of efficiency. These derive from two primary sources. The first is the existing instrument, which is time-consuming and laborious to apply, and is conducted outside the learning process. The second is the scope of the training and training maintenance required for effective application. Both difficulties derive in part from the roots of DA's linguistic assessment components in the field of testing.

In order to achieve DLIFLC's vision of providing all students with multiple iterations of PBFA, fully integrated within the learning process with IBFA, re-design of DA is required. The goal would be a continuous dynamically oriented assessment of each learner, establishing a baseline learner profile and linguistic profile, then tracking these through observation, "re-setting" the LP periodically based on subsequent DA iterations.

A more distant goal is the full integration of formative assessment into the learning process through IBFA alone, by instructors who possess the skills to apply DA's components spontaneously and intuitively in day-to-day instructional activities. In the real world, that goal is difficult to attain. In college/university contexts, L2 instructors often have backgrounds and ongoing activities in other fields. In the DLIFLC context as well, alongside instructors with degrees preparing them specifically for L2 instruction, many others bring valuable experiences in other fields to their positions, receiving cross-training at DLIFLC to become effective language instructors. In such an environment and given DLIFLC's need to process thousands of learners simultaneously year-round with accountability for attainment of proficiency goals and requirements, it is unlikely that the "interventionist" PBFA approach can be entirely superseded. The practical goal must therefore be a maximally effective and efficient combination of PBFA and IBFA through a streamlined DA.

References

Aljaafreh, A., and Lantolf, J.P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *Modern Language Journal 78*, 465–483. https://doi.org/10.1111/j.1540-4781.1994.tb02064.x

Bernhardt, E.B., & James, C.J. (1987). The teaching and testing of comprehension in foreign language learning. In D.W. Birckbichler (Ed.), *Proficiency, policy and professionalism in foreign language education* (pp. 65–81). Lincolnwood: National Textbook Company.

Child, J.R. (1987). Language proficiency levels and the typology of texts. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations, and concepts* (pp. 97–106). Lincolnwood: National Textbook Company.

Coalition of Distinguished Language Centers (CDLC). (2008). *What works: Helping students reach native-like second-language competence*. Hollister: MSI Press.

Corin, A.R. & Leaver, B.L. (forthcoming). *Fields of the Mind: History, Theory, & Application of Cognitive Field Concepts to Language Learning.* Hollister: MSI Press.

Davis, R.L. (n.d.). Reading performance tasks. (http://pages.uoregon.edu/rldavis/readingtasks/whyperformance.html).

Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *Modern Language Journal 78*(3), 273–284.

Dörnyei, Z., & Csizér, K. (1998). Ten commandments for motivating language learners: Results of an empirical study. *Language Teaching Research 2*(3), 203–229.

Ehrman, M. (1996a). Learners and teachers: The application of psychology to second language acquisition. *Russian Language Journal* 55, 33–71.

Ehrman, M. (1996b). *Understanding second language learning difficulties*. Thousand Oaks: Sage.

Ehrman, M. (1999). Ego boundaries and tolerance of ambiguity in second language learning. In J. Arnold (Ed.), *Affect in language learning* (pp. 68–86). Cambridge: Cambridge University Press.

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

Ehrman, M., & Leaver, B.L. (2002). *E&L learning style questionnaire v. 2.0*. (2002). http://www.cambridge.org/resources/0521837510/2127_Leaver%20learning%20styles%20test.DOC

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

Fleming, N.D., & Mills, C. (1992). Not another inventory, rather a catalyst for reflection. *To Improve the Academy 11*, 137–149.

Lantolf, J.P., & Thorne, S.L. (2006). Sociocultural theory and second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 201–224). Mahwah: Erlbaum.

Leaver, B.L. (1997). *Teaching the whole class*. Thousand Oaks: Corwin.

Leaver, B.L., Ehrman, M., & Lekic, M. (2004). Distinguished-level learning online: Support materials from LangNet and RusNet. *Foreign Language Annals 37*(4), 556–565.

Leaver, B.L., Ehrman, M., & Shekhtman, B. (2005). *Achieving success in second language acquisition*. Cambridge: Cambridge University Press.

Leaver, B.L., & Oxford, R. (2001-2005). Individual difference theory in faculty development: What faculty developers should know about style. *Russian Language Journal 55*, 73–127.

Pintrich, P.R., Smith, D.A.F, Garcia, T., & McKeachie, W.J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. National Center for Research to Improve Post-Secondary Teaching and Learning (U. of Michigan), http://files.eric.ed.gov/fulltext/ED338122.pdf

Poehner, M.E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 Development*. Springer.

Stoffa, R., Kush, J.C., & Heo, M. (2011). Using the Motivated Strategies for Learning Questionnaire and the Strategy Inventory for Language Learning in assessing motivation and learning strategies of generation 1.5 Korean immigrant students. *Education Research International* 2011, 8 pages. Article ID 491276

Wigglesworth, G. (2008). Task and performance based assessment. In Elana Shohamy (Ed.), *Language Testing and Assessment* (2nd ed., pp. 111–122). Springer.

About the Authors

Andrew R. Corin is Professor Emeritus at the Defense Language Institute Foreign Language Center. He is former Dean of DLIFLC's School of Resident Post-Basic Education, also former Dean of Educational Support Services for Post-Basic Education. His primary languages of teaching are Serbian-Croatian and Russian.

Currently a military interpreter of Ukrainian, stationed abroad, Sergey Entis is Academic Specialist (retired) at the Defense Language Institute Foreign Language Center, where he made contributions for several years as the Diagnostic Assessment Center Director and as Senior Academic Advisor to the Provost. His primary languages of teaching are Russian and Ukrainian.