Defense Language Proficiency Testing System 5 Framework

(DRAFT)

Test Development Division
Evaluation and Standardization
Defense Language Institute Foreign Language Center

Contact Information:
Mika Hoffman, Dean
Test Development Division
Evaluation and Standardization
Defense Language Institute Foreign Language Center
Mika.Hoffman@us.army.mil
(831) 242-5553

Table of Contents

1. Purpose

The purpose of the DLPT5 Framework is to provide stakeholders and prospective test users with information regarding the purpose of the Defense Language Proficiency Test 5 (DLPT5) Testing System, its test design, and how the Interagency Language Roundtable (ILR) Language Skill Level Descriptions are used as the basis of test constructs, test development, test use, and test score interpretation and decision-making based on scores. Section 2 provides a description of what the DLPT5 testing system is. Section 3 describes the theoretical underpinnings of the test constructs: reading and listening comprehension ability and the sub-abilities DLPT5 tests measure in relation to the ILR. Section 4 describes the process of generating the DLPT5 operational test forms. Section 5 outlines plans in support of the DLPT5 testing system, and finally, Section 6 describes future plans to make the testing system more efficient and user-friendly. It is our hope that this framework document will help stakeholders and test users better understand the nature of the DLPT5, the issues relating to the use of the DLPT5 testing system, and the system's capabilities and limitations, so that they will be able to make informed decisions about using DLPT5 test scores. The Test Development Division at the Defense Language Institute Foreign Language Center (DLIFLC) was tasked to develop this document by the Defense Language Office (DLO). The initial drafts have been developed within DLIFLC with input from the Defense Language Testing Advisory Board (DELTAB).

2. Overview of the DLPT5 Testing System

The DLPT5 Testing System is designed to assess the global language proficiency in reading and listening of native speakers of English who have learned a foreign language as a second language and speakers of other languages with very strong English skills. The DLPT5 tests measure proficiency as defined by the ILR Skill Level Descriptions, levels 0+ – 4 (see Appendix A) and are used to document and make operational readiness, incentive pay, assignment and training decisions for civilian and military personnel with language skills in the United States government. All DLPT5s are delivered on computer.

2.1 Test design

DLPT5s in many languages include both a lower-range test and an upper-range test. The lower-range test measures ILR proficiency levels 0+ - 3, while the upper-range test measures ILR proficiency levels 3+ and 4. Examinees will normally take the lower-range DLPT5; those who receive a score of 3 on this test may be eligible to take the upper-range test, depending on the policy of their organization.

There are two test formats for the DLPT5: multiple-choice (MC) and constructed-response (CRT). The multiple choice format is used for languages with large examinee populations, and tests are scored by computer. The constructed response format is used for languages with small examinee populations, and tests are scored by human raters certified to score such tests. Multiple choice or constructed response formats may be used for both lower-range and upper-range tests. A description of DLPT5 in both the multiple choice and constructed response formats follows.

2.1.1. DLPT5 in Multiple-Choice Format:

Upper-Range:

- The Upper-Range Reading Test contains approximately 36 questions with approximately 14 authentic written passages. Each passage may have up to 5 questions with four answer choices per question.

- The Upper-Range Listening Test contains approximately 36 questions with approximately 14 authentic audio passages. Each passage may have up to 3 questions with four answer choices per question. All passages are played twice.

- For research purposes, some questions are not scored. These questions do not count toward the final score the examinee receives. Examinees are told that such questions are in the test but are not told which questions are the unscored ones.

Lower-Range:

- The Lower-Range Reading Test contains approximately 60 questions with approximately 36 authentic written passages. Each passage may have up to 4 questions with four answer choices per question.

- The Lower-Range Listening Test contains approximately 60 questions with approximately 37 authentic audio passages. Each passage may have up to 2 questions with four answer choices per question. Passages at the beginning of the test are played once. Starting from level 2, examinees hear the passages twice.

- For research purposes, some questions are not scored. These questions do not count toward the final score the examinee receives. Examinees are told that such questions are in the test but are not told which questions are the unscored ones.

2.1.2. DLPT5 in Constructed-Response Format:

Upper-Range:

- The Upper-Range Reading Tests contains approximately 35 questions with 12 authentic written passages. Each passage has two or three questions.

- The Upper-Range Listening Test contains approximately 35 questions with 12 authentic audio passages. Each passage has two or three questions and is played twice.

Lower-Range:

- The Lower-Range Reading Test contains 60 questions with 30 authentic written passages. Each passage may have up to 3 questions.

- The Lower-Range Listening Test contains 60 questions with 30 authentic audio passages. Each passage has two questions and is played twice.

2.2 Test Content

The DLPT5 is designed to measure proficiency in the target language regardless of how it has been acquired. For this reason, and because of the broad proficiency orientation of the test, its content is not tied to any particular language-training program.

The DLPT5 is designed to measure examinee ability to understand materials produced by native speakers of the target language for communicative purposes. Thus, the test passages need to reflect the characteristics and the types of written and spoken texts in the target-language-use domain, rather than being texts produced for instructional purposes. Whenever possible, the DLPT5 passages are sampled from real-life sources such as signs, newspapers, radio and television broadcasts, the Internet, etc. When such sampling is not possible, passages must nevertheless have the characteristics of written or spoken texts used in real life for communicative purposes, as judged by native speakers of the target language. The passages cover a broad range of content areas, including social, cultural, political, economic, geographic, scientific, and military topics. Table 2-1 provides a select description of the content and topics DLPT5 samples. It should be noted that Table 2-1 is by no means an exhaustive list of topics the DLPT5 tests cover, and the ratio of the Final Learning Objective (FLO) content areas in each of the ILR skills levels measured varies slightly from language to language. In order to maximize content authenticity within each language, the DLPT5 test specifications allow for language-to-language variation in the content covered, even though this may lead to somewhat less standardization across languages.

Table 2-1

FLO Content Area	**Example Sub-categories**
Level 1	
Military-Security	Military ranks Basic police subjects (arrests, etc.) Customs officials Traffic regulations
Economic-Political	Government ministries National events Hiring and promotion Marketplace activities Basic bank transactions Basic travel and tourism
Scientific and Technological	Health services (appointments, prescriptions, hospitals, etc.) Technological devices of daily life Simple texts on scientific discoveries / research

Cultural and Social	Family event announcements Holidays School events Cultural fairs Sports Entertainment Obituaries
Geography: Physical, Political, Economic	Landmarks and spatial orientation Weather and climate Basic geographic relations
Level 2	
Military-Security	Warfare activities Military career and training Casualty reports Arms sales and disarmament Weaponry and equipment Law enforcement issues Crime and violence Terrorism Smuggling Military intelligence issues
Economic-Political	Transportation and travel Population trends Agriculture issues Trade issues (WTO, tariffs, export/import) Financial issues (budgets, inflation, taxes) Economic growth Industrialization Employment Political systems Elections
Scientific and Technological	Medical research and trends Medical / scientific training Inventions and new treatments Discoveries Technological progress
Cultural and Social	Family issues (marriage, divorce, etc.) Women's rights / status Educational issues Customs and traditions Religious subjects Leisure, art, and entertainment activities Media issues (freedom of the press, etc.)

Geography: Physical, Political, Economic	Climate issues Topography Urban vs. rural areas Historical / famous sites Natural disasters Pollution and other environmental issues Water issues
Level 3	
Military-Security	Analysis and discussion of military events and intelligence activities Assessment of military cooperation or alliance among countries Rationales for arms sales or military aid Description of advanced weapons systems Analysis of historical battles Analysis of security issues: justifications for security measures, etc.
Economic-Political	Economic analysis Political analysis Economic policies Assessment of a country's economic situation Development and growth issues Political strategies Political reforms and democracy Analysis of international political and trade relations
Scientific and Technological	Explanation of scientific, medical, and technological issues Analysis of scientific achievements, discoveries, or findings (e.g., the ethics of cloning) Research policies of governments or institutions Impact of scientific progress on families, behavior, society, etc. Scientific theories

Cultural and Social	Evaluation or analysis of social events and relationships Discussion of societal issues (discrimination, violence, drugs, privacy, etc.) Private vs. public education, the role of government in education, standards, etc. Religious issues: the relationship of church and state, fundamentalism, persecution, value systems Freedom of the press Commentary on books, movies, artworks
Geography: Physical, Political, Economic	Analysis of environmental policies Opinions on water issues (dam-building, etc.) Analysis of border disputes Use of natural resources Plans for maintaining habitat (e.g., rainforest preservation, anti-desertification plans, etc.)
Level 4	"Think-pieces" pertaining to any topics related to FLO Written or spoken discourse in the vernacular or exhibiting individual characteristics in writing or speaking styles

2.3 Test Format

The DLPT5 is a bilingual test: The test passages are in the target language, but the rest of the test, including test instructions and test questions, is in English. All tests include instructions on how to take the test, examples of how to answer the questions, and question sets containing the following parts that examinees see on the computer screen:

- **Orientation**: This is a short statement in English that appears before each passage. Its purpose is to identify the context from which the passage was taken.

- **Passage**: This is the only element of the test that is in the target language being tested. The maximum length of a listening comprehension passage in the test is approximately 2.5 minutes. The maximum length of a reading comprehension

passage is approximately 400 words. Most of the passages are much shorter than the maximum length. Examinees see written passages on the computer screen for the reading test but listen to audio passages though headsets for the listening test.

- **Question statement**: Each individual question is based on the passage, is written in English, and is posed in the form of a complete question or an incomplete statement. The questions may ask about what is explicitly stated in the passage or, in some cases, what is implied in it. Occasionally questions may ask about expressions that are used in the passage.

- **Answer choices (in MC tests)**: Each question statement is followed by 4 answer choices, also written in English, only one of which is the correct answer to the question. Each answer choice is displayed on the screen with a button next to it that examinees will click to select that choice. Note-taking is not permitted, but examinees can change their selection by clicking on a different button.

- **Answer box (in CRT tests)**: For each question, there is a box on the screen in which examinees type their answer in English. Examinees may also type notes in these boxes if they wish.

2.4 Test Administration

DLPT5 has two separate tests, a Reading Test and a Listening Test, which are administered at separately-scheduled sessions and delivered via computer. Examinees have three hours to complete the Reading Test and three hours to complete the Listening Test. Approximately halfway through each test, examinees will be given a 15-minute break. The break does not count toward the overall test time.

2.5 Examinees and test users

The DLPT5 examinees are U.S. military and government civilian personnel with varying degrees of proficiency in a foreign language. They may be native speakers of English or native speakers of other languages with very strong English skills. Their jobs require them to deal with written and/or spoken texts in a foreign language in many different contexts ranging from teaching to linguistic deciphering to content analysis. Some jobs require rapid text processing to obtain the gist or locate specific pieces of information while others require careful analysis at the textual level for instructional purposes. These people take the DLPT5 tests annually as required by their specific job.

DLPT5 test users include the DLPT5 examinees, their supervisors, and people in government agencies who are using DLPT5 test scores for decision-making in areas such as job placement, graduation requirements, eligibility for further training, or receiving foreign language proficiency pay (FLPP).

3. Defining the test constructs

The DLPT5 test construct is the ability to understand written and spoken input intended for the general public in terms of the target-language use domain as specified by the ILR. The DLPT5 measures use of receptive language skills in context, i.e., listening comprehension ability and reading comprehension ability in a given target language for a communicative purpose. The DLPT5 test design is informed by current theoretical understanding of second language attainment and/or proficiency levels, research findings on the influence of texts and question types on test performance, and most importantly, the Interagency Language Roundtable (ILR) Skill Level Descriptions, which serve as the standard against which examinee linguistic proficiency is measured. The two constructs being measured, reading comprehension ability and listening comprehension ability, can be seen as the ability of people to interact in certain ways with certain kinds of authentic written or spoken texts, and so measuring these abilities necessarily involves the interplay among three variables: examinees, authentic written or spoken texts, and tasks. This section explains how we define and operationalize these variables in the context of the DLPT5 so that readers will have a better understanding of how examinee performance is related to the target language use domain, and what is expected with regard to language ability at different ILR levels. This understanding, in turn, will allow test users to make informed inferences about examinee language ability.

An examinee's ability with regard to communicative language use can be regarded as consisting of two components: language knowledge and strategic competence (Bachman & Palmer, 1996). Language knowledge is further broken down into four components: grammatical knowledge, textual knowledge, functional or illocutionary knowledge, and sociolinguistic knowledge. Grammatical knowledge and textual knowledge are the functional capacity to use vocabulary, syntax, phonology (in spoken language) or graphology (in written language), cohesion, and rhetorical organization. Grammatical knowledge and textual knowledge are involved in control of the formal structures of the language in order to understand grammatically acceptable utterances or written sentences and the understanding of how these utterances or sentences are organized to form texts. Functional and sociolinguistic knowledge, often referred to as the pragmatic component of language use, enable examinees to interpret language appropriate for the particular language use setting and to relate the spoken or written texts to the communicative goals of the speaker or writer.

Strategic competence refers to the strategies or cognitive and metacognitive processes examinees use to interpret language effectively and to answer questions. This ability is necessary for examinees to determine the most effective means to achieve a communicative goal and execute those plans. These strategies include deciding what goal to pursue, what resources are needed to achieve the goal, and how to employ those resources to achieve the goal. Examinees use different strategies at different proficiency levels and their choices of courses of action influence test performance. For example, those at higher ability levels tend to make better use of both cognitive and metacognitive processes, because their control of the linguistic elements are automatic, which frees more cognitive capacity for higher-level processes, such as making inferences, monitoring their comprehension, and paying attention selectively, etc., whereas lower-ability language users tend to rely on grammatical cues for meanings and their control of

target-language elements is not yet automatic, forcing them to use up much of their cognitive capacity for linguistic processing and leaving little room for higher-level processing.

Examinee performance on the DLPT5 depends on the integration of their language ability in the target language, their individual attributes apart from this ability, and the strategies they use. The purpose of the test is to measure ability in the target language so it is important to minimize the influence of extraneous factors on test performance.

The examinees' other individual attributes are not part of their language ability but may affect their performance on the test. These attributes include individual characteristics such as age, gender, and first language, real-world knowledge (including cultural or topical knowledge related to the target language or knowledge in a specialized area), and "affective schemata," meaning examinees' affective responses to test passages and questions. Affective responses may influence the way examinees process the passages and respond to the test questions. Various measures are taken in order to minimize the potential impact these factors might have on test performance. For example, in selecting test passages, passages that require specialized outside knowledge to understand are excluded, as are passages that could likely be understood on the basis of outside knowledge without need for the appropriate linguistic knowledge. Topics are avoided that would unduly advantage or disadvantage any particular age group, gender, or ethnic or religious group. Topics are also avoided that could be expected to evoke a strong negative emotional reaction in a large proportion of DLPT5 examinees (who are primarily military personnel). In addition, test anxiety is minimized by providing examinees with extensive familiarization materials.

Although some strategic competencies, as illustrated above, might be considered to be relevant to the construct, test-taking strategies that are construct-irrelevant also exist (e.g., choosing an answer based on its length or not choosing an answer that includes certain key words, such as "all" or "never"), and the effect of the latter on test performance is also minimized to the extent possible. In the construction of questions, care is taken to minimize examinee ability to guess the correct answer based on insufficient linguistic ability (e.g., answer choices are similar to each other in terms of length and syntax). In addition, the familiarization materials mentioned above also help to reduce differences in test performance based on differences in familiarity with regard to the types of test passages and questions.

Research findings, to be discussed in sections 3.2 and 3.3, suggest that text types and question types may affect test difficulty and test performance. Care has been given to ensure that all DLPT5 test passages target the general public and are coherent. The types of questions appearing in the DLPT5 tests are listed in Tables 3.2 and 3.4. Detailed discussions of text types and question types in the DLPT5 context is provided in sections 3.2 and 3.3.

Another significant potential construct-irrelevant variable affecting test performance is the wording and language of the questions themselves. Shohamy (1984) found that multiple-choice and open-ended items in examinee's native language were easier than in the target language. She speculated that the wording of questions in the examinees' first language may provide some clues to the general meaning of the passage and therefore help students to guess the correct answer in multiple-choice items. All DLPT5 questions are in English, and the test is designed

for native speakers of English. Native speakers of other languages may be disadvantaged by having the questions in English; writers do, however, attempt to keep the English of the questions (and answer choices, for multiple-choice tests) as simple as possible, to help mitigate the effects of differences in English reading proficiency. Less frequently used vocabulary may be unfamiliar to examinees and could pose difficulty for examinees in processing the questions and/or associated answer choices. The DLPT5 test developers attempt to reduce lexical processing load by choosing high-frequency words and simple syntactic structures where possible.

The purpose of the DLPT5 is to make inferences about examinee target-language ability, as reflected in the ILR Skill Level Descriptions, based on samples of language use represented in a range of test passages and questions. The first step in developing the test is thus to define language ability and the characteristics of receptive language use. The next step is to operationalize the assessment of language ability by producing test passages that represent relevant characteristics of the target language and developing questions that target the specific language abilities of interest. We maintain the linkage between the test and the ILR by a series of ILR-based content reviews and analyses of the outcomes of test administrations.

It should be noted that there are some purposes for which the DLPT5 is **not** an appropriate test. The DLPT5 is not intended to measure target language proficiency in speaking or writing, nor is it intended to reflect examinees' job-related performance or ability to perform specific language-related tasks under special circumstance (e.g., reading or listening to a target language passage indefinitely many times with the aid of supplemental reference materials and information sources). In addition, the DLPT5 is not intended to be used as a tool to diagnose strengths, weaknesses, or progress in learning a target language.

The ILR Skill Level Descriptions form a functional scale that provides a standard that defines what examinees at different levels of language proficiency can or cannot do with the target language for communication in the target-language environment. They consist of general characterizations of language users at each of five levels of proficiency, the highest level being that of the highly educated native speaker[1]. While these descriptions serve as the standard against which examinee performance is to be measured, they are not adequate in and of themselves to determine test content fully. In developing test material, DLPT5 developers must therefore rely on several supplemental strategies.

This section has four sub-sections. Section 3.1 explains the use of the Interagency Language Roundtable Skill Level Descriptions in the definitions of required examinee ability and some strategies that are employed to supplement these descriptions in the development of test material. Sections 3.2 and 3.3 outline our definitions of the test constructs of reading comprehension and listening comprehension, respectively, and discuss the effects of examinee ability, passage, and questions on test performance. Section 3.4 provides the rationale for measuring test-taker performance on the basis of the interplay of the examinee, the test input and the tasks in the DLPT5.

[1] The term "native speaker" is a controversial one in the fields of second-language acquisition and language assessment. For our purposes, we follow the notion as conceived by the originators of the ILR: a person who has grown up using the language and who has been educated in the language.

3.1 The Interagency Language Roundtable Skill Level Descriptions

The ILR Skill Level Descriptions provide the basis for DLPT5 test development and test score interpretation. This section explains how features of the ILR Skill Level Descriptions are used in DLPT5 test development and how the descriptions are interpreted in order to define the scope of person ability measured by the DLPT5. This section also explains how the descriptions of texts found in the ILR, supplemented by guidelines based on text typology, are used for test passage selection. In addition, explanation is provided on how supplemental strategies that make use of inferences from the functional nature of the ILR scale are used in the selection of passages for the test and the development of tasks based on those passages.

The ILR Skill Level Descriptions consist of a set of descriptions that characterize performance typical of each level of proficiency from 0 (no proficiency) through 5 (that of a highly educated native speaker), with "plus" levels from 0+ through 4+. These descriptions provide target-language-user profiles in the form of "can do" and "can't do" statements. The ILR also provides some examples of the types of texts second-language users are able to understand at the respective ILR levels in the skill modalities of interest to DLPT5. In the context of assessing reading and listening comprehension abilities, these "can do" and "can't do" statements form a set of explicit and implicit standards used to evaluate the communicative skills second-language users have acquired when they deal with texts produced by native speakers of the target language in real-life communicative situations.

The ILR Skill Level Descriptions deal with general rather than job-related language skills and stress proficiency, not achievement (Lowe, 1986). The use of the ILR Skill Level Descriptions as the standard for DLPT5 development thus implies a focus on whether examinees are able to understand a given target language (TL) for given communicative purposes in communicative situations. Secondly, the ILR emphasizes consistent and sustained language ability (Lowe, 1986). The ILR Skill Level Descriptions provide generalized user guidelines as to what a typical second- or foreign-language user can or cannot do at any given ILR level; for example, at Reading Level 2, the reader is characterized as follows: "able to read with some misunderstandings straightforward, familiar, factual material, but in general insufficiently experienced with the language to draw inferences directly from the linguistic aspects of the text;" "typically able to answer factual questions about authentic texts…;" "can locate and understand the main ideas and details in material written for the general reader;" and "generally the prose that can be read by the individual is predominantly in straightforward/high-frequency sentence patterns." The "can do" statements suggest what a typical Level 2 reader is able to accomplish and provide explicit points of reference for defining the areas of ability most essential at this level. The "can't do" statements such as "inability to discern nuances and/or intentionally disguised meaning" (Reading Level 2+), and "can get the gist of more sophisticated texts, but may be unable to detect or understand subtlety and nuance" (Reading Level 3), on the other hand, function as implicit references to the limit of what a typical reader is able to do. In addition, the ILR also acknowledges that language users have the ability to process language minimally at a higher level, e.g., at Reading Level 2+, "the individual is able to get the gist of main and subsidiary ideas in texts which could only be read thoroughly by persons with much higher proficiencies." In the context of testing, the DLPT5 only measures what examinees can do, and the assignment of ILR levels is based on the success rate of answering questions relating to the

"can-do" statements correctly. The "can't-do" statements provide explicit limitations on passage selection and task development at given levels. For example, Level 2 listeners are "able to understand facts; i.e., the lines but not between or beyond the lines." Therefore, tasks requiring listeners to make speaker-intended inferences are not included for Level 2 test passages.

When measuring the consistent and sustained aspect of language use ability, DLPT5 test developers focus on the quantity and quality of the examinees' language performance through tasks designed to elicit the sub-skills considered important at a given ILR level. For example, at Reading Level 2, a typical reader "can locate and understand the main ideas and details in material written for the general reader," and is "typically able to answer factual questions about authentic texts…" Questions at Level 2 thus focus on main ideas and supporting information in authentic texts that are straightforward in conveying factual information. The quantitative aspect of sustained performance is represented by the formulas used to assign scores on the DLPT5 (see section 3.4). The ILR also contains statements related to quality, or the accuracy with which target-language users are able to employ their ability. Level 2 readers read, for example, "with some misunderstandings," This qualitative aspect of performance is represented by the specific ideas required in selecting or producing the correct answer to a test question. Note that individuals with the same level of proficiency vary considerably in terms of both the qualitative and quantitative factors. It should be stressed that the DLPT5 measures general proficiency; two individuals who receive the same score on the test will often have different strengths and weaknesses.

The ILR Skill Level Descriptions anecdotally provide some descriptions of texts second-language users are able to understand; for example, in Reading Level 3, individuals are "able to read … with almost complete comprehension a variety of authentic prose material on unfamiliar subjects" and the text types include "news stories similar to wire service reports or international news items in major periodicals, routine correspondence, general reports, and technical material in his/her professional field…" The latter statement describes the types and nature of texts a typical reader at Level 3 is able to handle. However, a systematic treatment of both the nature and complexity of texts and their content is missing, which poses a problem for DLPT5 test developers when selecting reading and listening test passages. Child's text typology, discussed below, provides a more detailed treatment of texts in spoken and written forms at all ILR levels, and is therefore used by DLPT5 test developers as supplementary guidelines for text selection. In short, the skill level statements in the ILR provide descriptions of what language users can or cannot do over a wide spectrum of communicative situations, whereas the textual modes delineate the texts produced by native speakers for various communicative purposes.

Child's text typology (Child, 1987, 1998 & 1999) is a scale in the textual dimension for both written and spoken texts. DLPT5 test developers use Child's analyses and descriptions of texts found in the ILR Skill Level Descriptions as standards for selecting test passages. The typology of texts posits four text modes or purposes to express the ways in which language is used by native speakers for communicative purposes. The four modes in the text typology, presented as a "scheme of increasing complexity" are the Orientation Mode, Instructive Mode, Evaluative Mode, and Projective Mode, corresponding roughly to the ILR Skill Level Descriptions Level 1, Level 2, Level 3 and Level 4. Thoughts on textual modes from other ILR scholars such as Lowe

(1988, 1998), Clifford (1980), and Edwards on written texts (1996) are included in the current collection of textual analyses used by the DLPT5 test developers.

The most prominent characteristics of these textual modes, taken from Child's papers (1987, 1998, 1999), are provided below. Refer to Tables 3-2 and 3-4 for more details of textual analyses, descriptions of the texts considered appropriate at targeted ILR levels, and their sources. It should be noted that native speakers do not have textual mode in mind when they speak or write, and, therefore, texts do not usually fall under one single mode. A textual mode is determined by how dominant discourse and linguistic elements in a given mode are in a text.

- Orientation Mode

The purpose of texts covered by this mode is to make information immediately available in the form of simple spoken or written texts to those needing it. Texts in the Orientation Mode are usually concerned with orienting readers or listeners regarding who or what is where, or what is happening or supposed to happen within a generally predetermined pattern. Such texts are easily accessible to users who have acquired the basic second-language functions and elements because such texts are usually short and grammatically simple, yet they are "part of the real world" and convey communicative purposes as considered by native speakers of the target language. Some examples of texts in the Orientation Mode are indications of identity, arrivals and departures of mass-transit carriers, times and places of meetings, and, in listening contexts, greetings or other formulaic exchanges in daily conversation.

- Instructive Mode

The purpose of texts in the Instructive Mode, like the Orientation Mode, is to transmit factual information. However, the variety and scope of texts are much more expanded. Generally speaking, texts in this mode convey written or spoken information about something that "exists or is developing or should take place in the real world" but do not offer "analytical or intuitive judgments" concerning the information conveyed. Writers and speakers who produce such texts assume that the reader or non-participatory listener shares a sufficient amount of background information. In other words, topics in the Instructive Mode are either intrinsically familiar to users or they can be worked out from the content. Although texts in the Instructive Mode deal with factuality and follow a predictable discourse pattern, speakers or writers may use a variety of expressions to achieve the same communicative purpose. Second-language users must acquire the basic rules of the target language to comprehend texts in the Instructive Mode. Some examples of texts in the Instructive Mode include accounts of domestic and international events, detailed instructions on the assembly of a complicated piece of equipment, straightforward historical narrative, and a topographical description of a geographical area.

- Evaluative Mode

In texts in the Evaluative Mode, the emphasis on the transmission of factual information is shifted to "a perspective in which facts are selected and pressed into service in order to develop points of view; explain or apologize for personal conduct; state and defend past or projected policies." The speakers or writers of Evaluative texts intend the products to fulfill a social purpose. Texts in the Evaluative Mode contain analysis and evaluation of things and events about which both the speaker or writer and the intended audience share background information. Texts in the Evaluative Mode correspond roughly to ILR Level 3, a level at which second-

language users are expected to respond intellectually, intuitively, or instinctively to texts. In other words, users are expected to think in the target language when they process texts in the Evaluative Mode. Users must have good command of the target language and have developed a good deal of cultural knowledge of the target language. Some examples of Evaluative texts are newspaper editorials, radio commentaries, biography with some critical interpretation, personal correspondences providing justifications for past actions, and assessment of given policies.

- Projective Mode

Texts in the Projective Mode are the most difficult type to access due to their relative lack of shared information and assumptions on the part of the speaker or writer. The function of these texts is to take a novel or creative approach in order to rethink and verbalize solutions to problems either previously treated in a different way or not addressed at all. The speakers or writers are breaking new ground; thus the products are unique in their conceptualization and notable for their individuation. Such texts are relatively inaccessible because they reflect unfamiliar cultural values, highly idiosyncratic language use, or a combination of the two. Second-language users must be well acculturated into the target language society. They also must be sensitive to the ways the language used in Projective texts follows or diverges from the assumed linguistic or cultural norms. Some examples of texts in the Projective Mode are "think-pieces on the op-ed page of a newspaper or in a journal on the need to reformulate social, economic, or political policy on X, Y, or Z," and "exchanges between thoughtful people as they address, or reconsider personal problems or goals; discuss merits of a work of art or a performance from personal perspectives; or merely give voice to private feelings."

- Mixed mode

Mixed-mode texts refer to texts that display textual elements found in two modes. They are roughly equivalent to "plus" level texts; for example, if a text exhibits a blend of language elements from the Orientation Mode and Instructive Mode, this text will be treated as a Mixed-Mode (Orientation/Instructive) text. Such texts are roughly equivalent to 1+ texts. Standards for determining 1+ texts, according to Child (1999), should be the standards set for the Instructive Mode. By the same token, a Mixed-Mode Instructive/Evaluative text, i.e., a 2+ text, exhibits almost all characteristics found in the Evaluative Mode. It may fall short of the amount or extent of arguments, hypotheses, or evaluation generally found in a typical Evaluative text. In short, mixed-mode texts fall short of the characteristics required for the next mode on the typology hierarchy. Refer to Tables 3-2 and 3-4 for detailed descriptions of the characteristics associated with written and spoken texts at each proficiency level.

Standards in text typology apply to all languages in terms of text content (e.g., advertisement, public announcement, news story, editorial, etc.), function (e.g., to sell something, to attract people to an event, to inform about a recent happening, to offer an opinion), and linguistic complexity (e.g., range of lexicon, syntactic structures used, organizational characteristics of the text). Because the DLPT5 measures proficiency in a wide variety of target languages, selecting representative samples from the target language use domain is the first step towards ensuring test validity in the languages tested. Therefore, DLPT5 test developers map test passages against the text typology standards in terms of the content, function, and linguistic complexity factors. (These factors are expanded on in Tables 3.2 and 3.4) However, each target language tested has its own unique language features which must be accommodated. Examinees need to understand

the language and cultural elements in the target language they have acquired in order to process test passages associated with the different textual modes.

The ILR Skill Level Descriptions and Child's text typology provide two different perspectives on the ILR level of texts, and test developers make use of these perspectives when selecting texts to be used on the DLPT5. Child's text typology is a text-based perspective that directly addresses the level of a text based on its features. The ILR Skill Level Descriptions, on the other hand, is a language-user-based perspective that mentions selected text types at various levels as examples of the types of texts language users at that level can work with. The ILR skill descriptors were developed from observations of second-language user behavior. The textual descriptions, on the other hand, were derived from "the perspectives of the native speakers of the target languages" (Child, 1999). Our approach to testing examinee proficiency suggests that first, DLPT5 test developers need to select authentic passages used for real communicative purposes according to the textual standards prescribed and then they need to develop tasks that tap the abilities relevant to the level of the text in order to measure the examinee's language use ability described in the ILR Skill Level Descriptions. The use of these two scales has the further implication that examinees considered at a given level of proficiency are supposed to be able to demonstrate the sub-skills required at all lower levels of proficiency. For example, a reader who is at Level 2 in a target-language is supposed to be able to meet the criteria for levels 1 and 1+ as well.

As noted above, however, even the ILR Skill Level Descriptions and Child's text typology together are inadequate to determine fully the test passages selected, the sub-skills to be tested, and the tasks developed, and we therefore employ two supplementary strategies whose primary purpose is to describe the kinds of tasks language users can perform and the kinds of written or spoken texts they can comprehend or produce. First, we can determine what constitutes major or important information in a text at a given level based on what the practical effect would be of failing to understand that information. For example, in a level 1 spoken text about a change in tracks for a train departure, any information that is necessary in order to ensure a traveler would get on the right train would be considered important. Second, adhering to a practice followed by ILR experts, we infer that language-use tasks at a given level have functional equivalents across skill modalities; hence, descriptions of ability, accuracy, and text purpose applicable to one skill can be applied to another skill (see Lowe, 2006). For example, the skill level descriptions for level 2 in listening do not provide information about non-participatory listening text types. The level 2 reading skill level descriptions, however, refer to "descriptions and narrations in contexts such as news items describing frequently occurring events, simple biographical information, and simple technical material written for the general reader." We apply this description to functionally equivalent tasks in listening and infer that level 2 listening ability involves being able to comprehend descriptions and narrations in news broadcasts describing frequently occurring events, simple biographical information, and simple technical material for the general public.

The ILR Skill Level Descriptions also sometimes refer to characteristics of second-language users that we cannot practically test on the DLPT5. For example, the ILR Skill Level Descriptions from time to time describe second-language users' ability to process texts related to their areas of specialty, i.e., background knowledge; for example, in Reading Level 2, "persons who have professional knowledge of a subject may be able to summarize or perform sorting and

locating tasks with written texts that are well beyond their general proficiency level." In selecting test passages, it is difficult to avoid all subject matters that could be an area of specialized knowledge for an individual examinee. We do not think background knowledge should play a more major role in examinees' test performance than their language knowledge, but we acknowledge that background knowledge influences comprehension. It is our practice to control for examinees' background knowledge as much as possible by including a wide variety of text types and content representing target-language use situations for the general public in order not to favor particular groups of examinees.

The ILR is a description of a linear progression of proficiency that helps test users interpret test scores; however, it should be noted that language development is multidimensional and there is no clear indication of how sub-skills suggested in the ILR levels are ordered. Unless empirically verified, the descriptors are observations of second-language users' language-use ability in a wide spectrum of communicative situations. The skill descriptors are indicative of what a given examinee most likely is able to accomplish and where this examinee stands in the proficiency continuum.

3.2 Reading

In this section, the operational definition of the construct of reading comprehension ability will be provided. This section includes three sub-sections: The first offers the DLPT5 definition of reading comprehension. The second discusses factors affecting examinee performance to include examinee characteristics and text and question characteristics. Finally, the interplay of these three types of factors with ILR level will be described and presented.

3.2.1 What is reading?

Reading comprehension can be viewed from two perspectives: process and product (Alderson 2000). The process perspective looks at the interaction between the reader and the text, whereas the product perspective focuses on comprehension, the product of reading. During the process of reading, readers look at the text, decipher the symbols contained in the text, and decide what those symbols mean as well as how they are related to one another. Readers also think about all or some of the following: what they are reading, what the text means to them, how they can relate to other things they read or already know, and what they expect to come next. Readers may also be thinking if the text is useful, interesting or boring. Furthermore, they may be reflecting unconsciously about the difficulty or ease of the text. In short, components of language use are activated during reading, and the entire process is dynamic, and varies to readers of the same text at different times of reading and with different purposes. The purpose of reading also affects text processing and comprehension. The reader's purpose may be as varied as reading to identify specific information, reading for general comprehension, reading for instructional purposes, and/or reading to integrate information.

A common view of the reading process suggests that reading consists of two components: decoding and comprehension (Gough et al., 1992). Decoding consists of word recognition fluency and accuracy of word representation; the comprehension process refers to parsing sentences, understanding sentences in discourse, building a discourse structure, and integrating understanding of the text with what one already knows. The comprehension process describes both linguistic skills and cognitive and metacognitive skills. This view of reading argues that

fluency and accuracy of word recognition, reading rate, and processing efficiency are fundamental to reading ability because the comprehension process of parsing sentences and building a discourse structure requires a certain level of rapid and automatic processing: if a reader must keep most word meanings in conscious short-term memory, the reader's short-term memory capacity will be exhausted, and it will be very difficult for that reader to relate larger chunks of text to each other.

Carver (1992) further distinguishes five types of reading processes with distinct purposes: memorizing, skimming, scanning, learning and "rauding." Rauding is the normal reading in which the reader is comprehending most of the thoughts the author intends to convey, i.e., comprehension of the main ideas in the text. Learning, or reading to learn, involves constructing an organized representation of the text that includes major points and supporting details. The purpose of skimming and scanning is to locate discrete pieces of information in the text.

Reading can also be defined by its product, comprehension as demonstrated through the tasks that readers are able to accomplish. Ultimately, it is this product that is measured, not the process that produces the product. The reading comprehension product defined by the ILR and the DLPT5 involves a set of text and task variables reflecting what readers encounter in real-life reading situations. These variables include the length and organizational characteristics of the texts to be processed, the complexity of the syntax, the usage and frequency of the vocabulary found in the texts, and the complexity of the tasks the examinees must accomplish. These variables influence text and task difficulty, which in turn, affect test performance.

The DLPT5 Reading Test assesses reading from the following purpose perspectives: reading to find information, reading to understand main ideas or main points, and reading requiring examinees to integrate and connect detailed information conveyed by the author. Additionally, the DLPT5 defines the construct of reading comprehension ability as the examinees' ability to deal with texts written for the general public, as demonstrated by their ability to answer questions targeting specific sub-skills. The next section discusses ability, texts, and questions in more detail.

3.2.2 Defining reading comprehension ability

We consider reading comprehension in a testing situation to be the result of the interplay of examinee ability, text difficulty, and question difficulty. The first step toward accurate measurement of examinees' reading comprehension ability is to define the scope of the reader ability the DLPT5 test attempts to measure and the types of reading texts that readers must deal with. Next, examinees must demonstrate ability by performing tasks that assess the sub-skills relevant to the construct. Therefore, it is necessary to define the types of tasks to which examinees must respond.

The ILR Reading skill level descriptors set the general standards for what successful reading is, and these standards are further specified by using text modes, inferences from the skill level descriptions about which functions are most important at a given level, and functional inferences from the skill level descriptions for other skills. For example, at Reading Level 1, the ILR skill level descriptions for reading state that the reader "can read … simple language containing only

the highest frequency structural patterns and vocabulary" and "can get some main ideas." The skill level descriptions for speaking say, "the individual can typically satisfy predictable, simple, personal and accommodation needs." These descriptors suggest, adding inferences from Child's modes and functional considerations, that Level 1 readers are generally capable of getting main ideas and identifying specific information important for immediate needs. Hence these two sub-skills are listed as Level 1 requirement, and all questions at level 1 focus on these two sub-skills.

DLPT5 test developers follow the ILR and Child's text typology guidelines in the selection of texts. Test developers use the following criteria to characterize written texts: purpose/mode, text type (for example, announcements, news reports, editorials, letters), linguistic elements (i.e., lexical range and syntactic complexity), and text organization. Developers also attempt to ensure that the pool of passages selected represents a variety of each of these characteristics.

DLPT5 reading questions fall under two broad question types: literal comprehension, in which information is stated explicitly in the passage, and inferential comprehension, in which examinees are required to integrate information presented in the passage and to some extent, use their schemata to answer the questions.

Table 3-1 presents the elements considered essential in defining the construct and designing the reading test. Detailed descriptions of text characteristics, question types, and reading ability with its required sub-skills are given in Table 3-2.

Table 3-1

I. **Purpose**: to measure US civilian and military language analysts' reading comprehension ability in a foreign language

II. **Construct definition**: the extent of the ability as specified by the ILR to
- process samples of authentic written discourse automatically and reasonably quickly,
- understand the linguistic information that is unequivocally included in the text,
- make any author-intended inferences that are unambiguously implied by the content of the text.

III. **Characteristics of text and question**: presented in Table 3-2

IV. **Characteristics of examinees**: See Section 2.5

V. **Reading abilities**: required reading abilities are presented in Table 3-2 according to the ILR Skill Level Descriptions (Appendix A)

- Identifying examinee, text and question characteristics

Examinees

Examinee characteristics such as their knowledge of the world and the target-language culture, strategies they use to achieve a goal, knowledge about different text types, speed of word recognition, automaticity of reading processes, and eye movement, as well as physical characteristics such as gender, age, and personality, affect reading comprehension ability. All

these factors affect how readers process a text and the degree to which they understand the text. Study findings in first-language and second-language differences also provide some insight to second-language reading: second-language readers transfer their first language reading skills and strategies; the linguistic distance between the first language and second language and between the languages' orthographic features affects reading performance; and second-language readers have to attain a linguistic threshold in the target language before their first-language reading skills are transferred (Enright et al. 2000). Knowledge of the target language is the factor that most strongly affects readers' comprehension ability. However, the research findings suggest that many other factors contribute to their test performance. In selecting texts and tasks for the DLPT5, it is important to minimize factors that developers believe are not strongly related to language ability, for example background knowledge, gender, and age. We do this by presenting a variety of types of authentic texts on different content areas, and by structuring the questions so that they cannot be answered correctly based on background knowledge or logic alone.

Texts

Aspects of the text may facilitate or make difficult the reading process. The topic or content of the text may have an impact on comprehension, particularly if it engages the examinee's background knowledge. Clapham (1996), in her study of content effect of the International English Language Testing System (IELTS) reading test, reports subject knowledge could facilitate comprehension for examinees who scored between 60% – 80% of her grammar test. Those who scored below 60% of the grammar test were not able to understand texts even in their subject discipline, whereas those who scored above 80% had little difficulty understanding texts outside of their areas of study. The implication is that subject knowledge may facilitate comprehension for at least some examinees; given that examinee subject knowledge is not a variable that can be controlled, the requirement for content variety in DLPT5 passages is important in order to minimize the impact of the variable. Another step taken to minimize subject-knowledge impact is to ensure that texts are taken from sources catering to the general reader. Hale (1988) examined performance on the Test of English as a Foreign Language (TOEFL) reading tests and reported students in the humanities/social sciences and biology/sciences did better on passages related to their own fields of study than on other passages. However, the effect was small because the TOEFL reading passages were taken from general readings intended for a general audience. The content of the DLPT5 Reading Test passages covers the five content areas stated in the Final Learning Objectives (FLO), which covers a wide range of topics (Section 2.2). Care has been taken to ensure texts selected are for the general audience, cover the five areas, and specialized knowledge is not necessary to understand the information presented in the passages.

Text organization is another factor that can affect the reading process and comprehension. Transparent and predictable organization of texts facilitates comprehension, at least at some levels of proficiency. Findings by Kobayashi (2002) indicated that high-proficiency examinees performed better on tasks targeted at overall understanding when the organization of the texts was clear than when it was not. For lower-level examinees, however, text organization had little effect on performance on global comprehension items. The findings support a claim that there is a proficiency threshold that must be crossed before examinees are able to use comprehension skills and knowledge in other areas in the examinees' long-term memory to answer global

questions. The implication for passage selection is that particularly at level 2 and higher, when text organization begins to be a factor within the ILR framework, there may be texts at the same level with varying complexity of structure and organization, and the ability being tested by global comprehension questions may be different for different types of texts: some texts will require considerable higher-level processing, such as synthesizing information or making inferences, whereas others might rely more on lower-level processing, such as straight linguistic decoding. Again, as always, the goal is to present a variety of texts so that no one type of learner will be particularly advantaged or disadvantaged.

The DLPT5 test developers use two broad categories to characterize written texts: discourse features concerning the purpose, text organization and characteristics, and text types; and grammatical features relating to syntax and vocabulary of the text (See Table 3-2 for details). The purpose involves text mode and communicative intent; text organization and characteristics include pragmatic and rhetorical features of the texts; and text types describe the nature of the texts, for example, advertisement, narrative, exposition, argumentation, persuasion or evaluation.

The contribution of grammatical and lexical knowledge to reading efficiency has been supported by many research findings. Knowledge of the target-language syntax contributes to more efficient processing of information. It also helps establish propositions and disambiguates lexical meanings. When the syntax is complex, more syntactic knowledge is required to process it, and more effort is required. Syntactic complexity may therefore slow reading and lead to lower reading efficiency. The ILR sometimes mentions syntactic complexity as a factor differentiating between levels, as when it refers to "the highest frequency structural patterns" for level 1, or "may rely primarily on lexical items as time indicators" for level 1+, or "may experience some difficulty with unusually complex structure" for level 3. However, there is a considerable variety of syntactic patterns among texts at the same level for level 2 and above; as with text organization, syntactic complexity can make some texts at a given level considerably more difficult than other texts at the same level, and variety of syntactic complexity among texts at a given level is desired.

One possible way to control for the variable of syntactic complexity is to simplify texts. However, the study in text simplification by Strother and Ulijin (1987, in Alderson 2000) suggested that simplifying syntax does not necessarily make texts more readable. Furthermore, simplifying the texts disauthenticates the texts, and the ability to read simplified texts is not necessarily generalizable to the ability to read genuine texts. DLPT5 test developers select texts with syntax appropriate for the ability range of examinees at the desired level. Texts that might otherwise fit the level but that have syntax too complex for the level are discarded rather than simplified.

Research findings in both first-language and second-language suggest vocabulary is the important contributor to reading comprehension (Read, 2000). A concept of vocabulary threshold was suggested for second language reading comprehension but the actual threshold level in terms of the number of words required is still a topic of debate.

In the DLPT5 context, vocabulary is not assessed directly. Lexical knowledge is assessed through comprehension of content. The ILR Skill Level Descriptions set expectations that

vocabulary knowledge itself is not always a crucial component of proficiency; for example, level 2 states, "The individual does not have a broad active vocabulary (that is, which he/she recognizes immediately on sight), but is able to use contextual and real-world cues to understand the text." The requirement that texts be written for the general reader and the text type examples in the ILR, taken together, tend to predict the sort of vocabulary appropriate for each level. Test developers select texts such that the difficulty of the vocabulary is consistent with the ILR level of the, and texts selected reflect authentic characteristics of the target language. Vocabulary difficulty for any individual is correlated with that individual's own learning history, so that what may be difficult for some examinees is easy for others. Again, variety in content areas is a key component to minimizing the interference of this kind of individual variation.

The length of the text is another aspect that may affect test performance. The finding by Engineer (1977, in Alderson 2000) suggests that longer texts may allow language testers to tap discourse processing abilities rather than merely relying on syntactic and lexical knowledge; this would imply that for higher proficiency levels, longer texts might be in some senses easier to process than shorter ones. Shorter texts, of course, typically require less time to process and less effort for short-term memory, so at lower levels it may be more important to keep the length short. The DLPT5 Reading Test measures comprehension of authentic reading materials at the discourse level. Most reading passages are taken from sources produced for real-world communicative purposes and have different text types and communicative intents. Text type and communicative intent may themselves affect length, so text length varies considerably among texts at the same level. For practical reasons, in that we do not want examinees to spend time reading material about which we do not ask questions, we keep reading passages as short as possible and edit extraneous material when necessary. Nevertheless, in designing the DLPT5, it was decided that passage length would have to be extended beyond the 120 word limit used in the previous test generation, DLPT IV, especially for levels 2 and above. At ILR levels 2 and above, ILR tasks require understanding relationships among parts of the text, and short texts do not supply enough material to test these tasks. As the ILR level goes up, maximum passage-length limits increase, allowing for the selection of test passages that mirror real-world reading materials, both in terms of their content and the processing required to comprehend them. The content of such passages is sufficient a) to allow for full explication of level-appropriate ideas (e.g., sequence of events and cause/effect about a factual occurrence at level 2 or argumentation with support about an abstract topic at level 3) and b) to support the multiple, level-appropriate questions required to test those ideas. Refer to Table 3-2 for specific length limits.

Finally, the degree of authenticity is considered one important aspect in the DLPT5 tests. In selecting written texts for use as test passages on the DLPT5, it is important to select texts that exhibit features of typical, current written texts that can be found in the target-language use situation. For this reason, wherever possible, DLPT5 test developers select fully authentic texts, i.e., those which are produced by users of the target language and which are intended to be read by other users of the target language in the target-language culture. There are, however, various circumstances in which texts must be used that are not fully authentic.

For example, the fully authentic text might exceed length constraints for test passages at a given level, or it might contain embedded off-topic material. In such a case, the fully authentic text might be abridged for use as a test passage, as long as the test passage remains natural and

coherent. Also, depending on the target language, material in certain content areas might be difficult or impossible to collect. The same might also be true for certain text types at certain levels. In order to achieve a desirable sampling of text types and subject areas at each level, it is sometimes necessary to purpose-write test passages. Occasionally, an isolated element of vocabulary or syntax will be in conflict with the overall level of the passage and crucial to comprehending the passage, or there will be a spelling or grammatical error in the original authentic text. In such cases, also, limited editing is permitted, although systematic simplification of a text is not.

In the end, we rely on the expertise of target-language experts who have extensive knowledge of the target language and culture and can judge whether a given test passage could appear in the appropriate context for its text type and be regarded as natural by members of the target-language use community.

Issues regarding the standard written form (i.e., standard language) in different language communities represent a particular challenge for DLPT5 test developers. The ILR offers no guidance about languages in which many standard dialects, scripts, or fonts are associated with one single language. Similarly there are no ability statements in the ILR for such situations. Serbian and Croatian, Hindi and Urdu, Uzbek, and Kurmanji Kurdish offer examples of language communities that may have a different perception of what constitutes "standard language" or "standard script." As concerns these languages, stakeholders have given DLPT5 test developers guidance on which dialect, script, or font is to be used and in some cases has given guidance to use multiple scripts in the same test.

Questions

Question types and the language used in the questions affect test difficulty and examinee performance. Kintsch and Yarbrough (1982) studied the interaction between texts and two types of tasks: macro-level processing tasks, which have to do with global understanding, and micro-level processing tasks, which have to do with local and phrase-by-phrase understanding. They found that the rhetorical organization of the text affected performance on macro-level tasks, but that performance on micro-level tasks was not affected by rhetorical organization. This finding is also supported by Kobayashi (2002). The implication here is that different questions relating to the same text can differ considerably in difficulty, if one is a micro-level task and the other is a macro-level task.

Pearson and Johnson (1978) identified three types of questions: textually explicit questions, textually implicit questions and scriptally implicit questions. Textually explicit questions are questions in which question information and the answer are found in the same sentence, textually implicit questions require examinees to combine information across sentences, and scriptally implicit questions require readers to integrate text information with their background knowledge, as the correct answers to the questions cannot be found in the text itself. Davey and Lasasso (1984, in Alderson 2000) reported that their textually explicit questions were significantly easier than their textually implicit questions. These findings are consistent with the ILR expectations that level 1 readers cannot integrate information well across sentences; that level 2 readers can combine information across sentences but not "read between the lines"; and that level 3 readers can understand points that are implied rather than stated explicitly in the text. Davey (1988, in

Alderson 2000) suggests that multiple-choice questions asking for implied information require more cognitive processing than do those asking for explicit information, and that performance on such items may be affected by test-taking or problem-solving abilities. DLPT5 writers attempt to mitigate these effects for multiple-choice questions by writing the answer choices such that examinees below the desired level will be unable to distinguish between answer choices based solely on logic.

The DLPT5 Reading test questions primarily target macro-level comprehension rather than discrete syntactic or vocabulary points in the passages, given that the ILR is a functional scale that does not typically specify expectations for specific knowledge of vocabulary or grammar. Questions target two types of comprehension: literal comprehension, referring to understanding of facts presented in the passages; and inferential comprehension, which requires examinees to integrate information and use their world knowledge to infer the author's intent. Within these two types of questions there are several sub-skills being measured, for example, ability to understand sequences of events, or ability to understand cause and effect, both of which test literal comprehension.

Davey and Lasasso (1984, in Alderson 2000) examined item and reader characteristics, and found that selected-response items (i.e., multiple-choice questions) were easier than constructed-response items when examinees could only read the text once. However, when examinees were allowed to look back at the text, there were no significant differences between selected-response items and constructed-response items. The finding was true for both textually explicit and textually implicit items. In DLPT5 reading tests, examinees may read the passage as often as they like, and questions are displayed simultaneously with the passages. Based on Davey and Lasasso's findings, then, we expect that whether a DLPT has multiple-choice or constructed-response questions do not have a significant effect on difficulty.

In sum, then, there are many factors that affect performance on reading tests, some that can legitimately be viewed as components of reading ability, such as automaticity of processing and vocabulary knowledge, and others that DLPT5 developers see as construct-irrelevant variables, such as background knowledge. To the extent possible, DLPT5 developers use a variety of texts in terms of content, purpose, and type, and a variety of task types, to cancel out individual examinees' advantages or disadvantages in particular construct-irrelevant areas.

- Describing reading abilities, texts and question types

Table 3-2 lists the ILR skills required of the examinees to demonstrate. It also describes texts and question types according to the ILR and text typology guidelines for score interpretation, and spells out the processing demand required of the examinees to perform successfully.

Under the heading of each ILR level, seven sub-headings are listed to describe ways in which examinee ability, texts, and questions are related according to the ILR. These seven sub-headings are:
ILR Skill Level Descriptions: targeted ILR skill level descriptions are provided verbatim,
Description of Expected Ability: the abilities considered to be typical at the given ILR level are provided,

Skills to be Assessed: salient abilities are operationalized in terms of measureable skills,
Target Language Input: under this heading, the purpose of reading, types of texts, and the characteristics associated with those texts are provided,
Focus of Task/Question: the types of questions are listed,
Sources: lists showing where test passages originate are given, and
Cognitive Load: the expected processing demands to be met by successful examinees are listed.

The content of the DLPT5 Reading Test includes areas and topics covered in the Final Learning Objectives (FLOs). Refer to Table 2-1 for select examples of possible content areas within the FLOs.

The channel of input (i.e., reading passages in the target language) is visual; that is, examinees read the target language input and questions. Please note that in the Reading Test, the specification for the maximum length of the reading passage is based on the rough English rendering of the target language text. The reason for this is to provide a rough equivalence across languages, since different target languages have different approaches to what constitutes a word, and the information conveyed in a certain number of words in English might be conveyed in many more words in one language, and many fewer in another.

Table 3-2

ILR Level	**Reading – Level 10**
ILR Skill Level Descriptions	Sufficient comprehension to read very simple connected written material in a form equivalent to usual printing or typescript. Can read either representations of familiar formulaic verbal exchanges or simple language containing only the highest frequency structural patterns and vocabulary, including shared international vocabulary items and cognates (when appropriate). Able to read and understand known language elements that have been recombined in new ways to achieve different meanings at a similar level of simplicity. Texts may include descriptions of persons, places or things: and explanations of geography and government such as those simplified for tourists. Some misunderstandings possible on simple texts. Can get some main ideas and locate prominent items of professional significance in more complex texts. Can identify general subject matter in some authentic texts. [Data Code 10]
Description of Expected Ability	Able to - comprehend very simple connected prose. - understand known language elements that have been recombined in new ways to achieve different meanings at a similar level of difficulty. - understand main ideas and/or simple, explicitly stated details. - identify general subject matter.
Skills to Be Assessed	Ability to understand - main ideas and/or general subject matter - explicitly stated, important information
Target Language	

Input	
Text Mode / Purpose	Orientation (Child) – Texts are intended to inform the reader as to the "who, what, when, or where" of events in his/her immediate surroundings, are of a predictable nature and fulfill basic social or practical functions.
Text Types	Texts are related to social or practical activities, e.g., personal invitations, requests for or offers of help, simplest instructions, bulletin board information, flyers, congratulatory messages, etc., or may be simple descriptions of persons, places, or things.
Text Organization and Characteristics	Texts at this level generally consist of loosely connected sentences, and reordering sentences within a given text does not affect meaning. Texts serve to fulfill a relatively restricted set of social functions, i.e., social survival and matters "of the moment," and information may quickly become irrelevant. Texts are simple and may be simplified on occasion or purpose-written, but they must exhibit the characteristics of the target language at this level: that is, they should be recognizable as plausibly authentic passages found in real-world sources.
Lexical Range	Most basic and most frequently used vocabulary (including shared international vocabulary and cognates) that relate to social and practical needs in daily life; vocabulary almost always conveys concrete rather than abstract notions. Vocabulary is usually generic rather than specific, e.g., "coat" rather than "windbreaker" or "parka."
Syntactic Complexity	Sentences represent the most basic or formulaic structures of the target language. They tend to be short and simple and are used primarily to refer to the present; compounding may sometimes occur.
Length	Up to 60 words according to the English rendering.
Sources	Classified ads, brochures, flyers, public announcements, bulletin board information, personal invitations, other simple personal or business correspondence; phone or "while-you-were-away" messages, tourist information, etc.; the aforementioned may come from standard print sources or from the Internet.
Focus of Task/Question	- Main idea or general subject matter - Explicitly stated, simple information
Cognitive Load	Examinees are required to process factual and explicitly stated information in the target language, locate the information, and identify the correct option in MC items or provide the correct answer for CRT items. The relationship between the questions with their expected responses and the target language input is direct as expected responses at this level are either literal translations or close paraphrases of the target texts.

ILR Level	**Reading – Level 16**
ILR Skill Level Descriptions	Sufficient comprehension to understand simple discourse in printed form for informative social purposes. Can read material such as announcements of public events, simple prose containing biographical information or narration of events, and straightforward newspaper headlines. Can guess at unfamiliar vocabulary if highly contextualized, but with difficulty in unfamiliar contexts. Can get some main ideas and locate routine information of professional significance in more complex texts. Can follow essential points of written discussion at an elementary level on topics in his/her special professional field. In commonly taught languages, the individual may not control the structure well. For example, basic grammatical relations are often misinterpreted, and temporal reference may rely primarily on lexical items as time indicators. Has some difficulty with the cohesive factors in discourse, such as matching pronouns with referents. May have to read materials several times for understanding. [Data Code 16]
Description of Ability	Able to - comprehend - understand known language elements that have been recombined in new ways to achieve different meanings at a similar level of difficulty. - understand main ideas and/or simple, explicitly stated important information
Skills to Be Assessed	Ability to understand - main ideas and/or general subject matter - explicitly stated, important information
Target Language Input	
Text Mode / Purpose	Mixed Orientation/Instructive Mode (Child) – Texts are intended to inform the reader as to the "who, what, when, or where:" of events in his/her immediate surroundings which are of a predictable nature and fulfill basic social or practical functions.
Text Types	Texts are similar to Level 10 texts but texts may also contain simple and straightforward biographical information or description.
Text Organization / Text Characteristics	Texts may show the structure typical of a paragraph, i.e., sentences connected in such a way that reordering of them is almost impossible. Texts serve to fulfill a relatively restricted set of social functions i.e., providing information for social survivals and matters "of the moment," and information may quickly become irrelevant. Texts are selected from the real world. Texts can also be minimally simplified to reflect the level

	of complexity required at this textual level.
Lexical Range	Most generic and most frequently used vocabulary (including shared international vocabulary and cognates) that relates to social and practical needs in daily life; vocabulary almost always convey concrete rather than abstract notions. Topic-specific words begin to appear within the context of basic social and practical situations.
Syntactic Complexity	Sentences representing the basic or more frequently used structures of the target language at this level. Compounding of verbs or the stringing of adverbial information begin to appear at this level. Verbs are still primarily used to refer to the present.
Length	Up to 90 words according to the English rendering.
Sources	Similar to Level 10 sources. But simple biographical materials may also be used at this level.
Focus of Task/Question	- Main idea and/or general subject matter - Explicitly stated, important information
Cognitive Load	Examinees are required to process factual and explicitly stated information presented in the target language, locate the information, and identify the correct option in MC items or provide the correct answer for CRT items. The relationship between the questions with their expected responses and the target language input is direct as the expected responses at this level are close paraphrases of the target texts.

ILR Level	**Reading – Level 20**
ILR Skill Level Descriptions	Sufficient comprehension to read simple, authentic written material in a form equivalent to usual printing or typescript on subjects within a familiar context. Able to read with some misunderstandings straightforward, familiar, factual material, but in general insufficiently experienced with the language to draw inferences directly from the linguistic aspects of the text. Can locate and understand the main ideas and details in material written for the general reader. However, persons who have professional knowledge of a subject may be able to summarize or perform sorting and locating tasks with written texts that are well beyond their general proficiency level. The individual can read uncomplicated, but authentic prose on familiar subjects that are normally presented in a predictable sequence which aids the reader in understanding. Texts may include descriptions and narrations in contexts such as news items describing frequently occurring events, simple biographical information, social notices, formulaic business letters, and simple technical material written for the general reader. Generally the prose that can be read by the individual is predominantly in straightforward/high-frequency sentence patterns. The individual does not have a broad active vocabulary (that is, which he/she recognizes immediately on sight), but is able to use contextual and real-world cues to understand the text. Characteristically, however, the individual is quite slow in performing such a process. Is typically able to answer factual questions about authentic texts of the types described above. [Data Code 20]
Description of Ability	Able to - comprehend straightforward, factual authentic material on concrete topics - understand main ideas and/or important information such as a major detail - follow the development of events described
Skills to Be Assessed	Ability to understand - main ideas - major details - sequence of events including cause and effect
Target Language Input	
Text Mode / Purpose	Instructive Mode (Child) – Texts are intended to inform the reader about factual information concerning concrete events or actions in the reader's world.
Text Types	Texts including the following: - narratives, e.g., news articles on familiar topics, events, biographies,

	obituaries - instructions / directions, - descriptions or elaborations with detailed descriptions of places, persons or objects or other physical phenomena - comparisons / contrasts
Text Organization / Text Characteristics	Texts are densely packed with factual (concrete, not abstract) information about situations and events. Texts are organized in a way that information is presented in a straightforward, predictable sequence and the presence of the author is not detected in the written material, i.e., the content is neutral / devoid of comment or interpretation on the author's part. Texts at this level are almost always taken from the real world but may, in rare instances, be simplified. Texts are intended for the general reader.
Lexical Range	Vocabulary used in this level is primarily concrete but topic-specific and may be substituted with synonyms. Vocabulary is sufficient to describe the basics about the concrete world and actions in it and goes beyond immediate survival needs. Vocabulary excludes most words related to the realm of the abstract.
Syntactic Complexity	Texts contain simple, compound, and/or complex sentences and may contain compound-complex sentences. Verbs reflect all timeframes. Features such as aspect, mood, and voice begin to appear. For inflectional languages, most nominal (case) forms are likely to appear in texts at this level.
Length	Up to 250 words according to the English rendering.
Sources	Newspapers, magazines, brochures, correspondence of a personal or business nature, or similar materials on the Internet.
Focus of Task/Question	Understanding - Main ideas - Important information and/or major details - Sequence of events - Cause and effect
Cognitive Load	Examinees do not "have a broad active vocabulary" but are able to "use contextual or real life cues to understand the texts." Examinees are expected to have some understanding of how the target language culture functions at a concrete level, i.e., they have basic target language cultural knowledge through which to interpret concrete information being read. Examinees are required to process factual information presented in the rhetorical structures characteristic of the text types at this level. Examinees are required to integrate and synthesize the factual information and identify the correct option in MC items or provide the correct answers in CRT items. The relationship between the target texts and its associated questions and expected responses are less transparent compared to Level 1+. Examinees are required to synthesize information from various parts of the text in order to answer questions.

ILR Level	Reading – Level 26
ILR Skill Level Descriptions	Sufficient comprehension to understand most factual material in non-technical prose as well as some discussions on concrete topics related to special professional interests. Is markedly more proficient at reading materials on a familiar topic. Is able to separate the main ideas and details from lesser ones and uses that distinction to advance understanding. The individual is able to use linguistic context and real-world knowledge to make sensible guesses about unfamiliar material. Has a broad active reading vocabulary. The individual is able to get the gist of main and subsidiary ideas in texts which could only be read thoroughly by persons with much higher proficiencies. Weaknesses include slowness, uncertainty, inability to discern nuance and/or intentionally disguised meaning. [Data Code 26]
Description of Ability	Able to - comprehend straightforward, factual authentic material on concrete topics - understand main ideas and major details - follow the development of the ideas - draw simple inferences based on information presented in the text - detect emotional overtones
Skills to Be Assessed	Ability to - understand main ideas, important information, major details, and sequence of events - draw simple, text-based inferences or conclusions
Target Language Input	
Text Mode / Purpose	Mixed Instructive/Evaluative Mode (Childe) – Texts are primarily intended to inform the reader about the factual information primarily concerning concrete events or actions in the reader's world. Texts may call upon the reader to draw simple conclusions or inferences based on factual information.
Text Types	Texts may include the following: - narratives, e.g., news articles, biographies, obituaries - instructions and directions - descriptions of persons, places, or things - explanations, e.g., expository writing
Text Organization / Text Characteristics	Texts are densely packed with factual (concrete, not abstract) information about situations and events. Information is not necessarily presented in a predictable or straightforward way. The treatment of concrete topics in such texts may occur in contexts or situations with which they are not normally associated. Texts may contain subtle choosing and ordering of

	material and of interpretative comment on the author's part; these choices give some evidence of the author's personal involvement in the material.
Lexical Range	Vocabulary is increasingly topic-specific and precise, and is sufficient to describe a wide range of fact-based actions and occurrences. Lexical Range begins to include vocabulary related to the realm of the abstract.
Syntactic Complexity	Texts frequently contain compound, complex sentences, and may also contain compound-complex sentences. A wide range of verbal forms such as tense, aspect, mood, voice and of nominal (case) forms may be used to discuss / describe the topics being written about.
Length	Up to 250 words according to the English rendering
Sources	Newspapers, magazines, books, Internet, brochures/pamphlets, correspondence of a personal or business nature.
Focus of Task/Question	- Understanding of main idea, major details, and sequence of events - Drawing simple inferences or conclusions
Cognitive Load	Though examinees exhibit weaknesses such as "slowness, uncertainty, inability to discern nuance and / or intentionally disguised meaning," they have "a broad active reading vocabulary" and "are able to use linguistic context and real-world knowledge to make sensible guesses about unfamiliar material." Examinees are expected to have a fairly well developed understanding of how the target language culture functions at a concrete level and an emerging understanding of the social-linguistic aspects of the target-language culture. Examinees are required to process information presented in the texts including the emotional overtones embedded. Examinees are required to understand and integrate the information, and identify the correct option in MC items or provide the correct answers in CRT items. The relationship between the target texts and its associated questions and expected responses are less transparent compared to Level 2. Examinees are required to synthesize information from various parts of the text in order to answer questions. On occasion, examinees are required to express judgments about the target texts in ways that are less directly related to the texts those judgments are based on.

ILR Level	**Reading - Level 30**
ILR Skill Level Descriptions	Able to read within a normal range of speed and with almost complete comprehension a variety of authentic prose material on unfamiliar subjects. Reading ability is not dependent on subject matter knowledge, although it is not expected that the individual can comprehend thoroughly subject matter which is highly dependent on cultural knowledge or which is outside his/her general experience and not accompanied by explanation. Text-types include news stories similar to wire service reports or international news items in major periodicals, routine correspondence, general reports, and technical material in his/her professional field; all of these may include hypothesis, argumentation and supported opinions. Misreading rare. Almost always able to interpret material correctly, relate ideas and "read between the lines," (that is, understand the writers' implicit intents in text of the above types). Can get the gist of more sophisticated texts, but may be unable to detect or understand subtlety and nuance. Rarely has to pause over or reread general vocabulary. However, may experience some difficulty with unusually complex structure and low frequency idioms. [Data Code 30]
Description of Ability	Able to - read a variety of prose material on unfamiliar subjects. - almost always interpret material correctly, relate ideas, and "read between the lines;" rarely misreads texts "at level."
Skills to Be Assessed	Ability to - understand main ideas / major or specific details - understand / detect implications or inferences intended by the author - understand supporting arguments - understand comparisons of points of views - understand vocabulary/idiomatic expressions in context - integrate and synthesize information/ideas and draw appropriate conclusions - identify the writer's tone / attitude / position - separate facts from opinions
Target Language Input	
Text Mode / Purpose	Evaluative Mode (Child) – Texts are written with a social/public purpose, i.e., to respond to facts, situations, or events with analysis, opinions, commentary, or feedback.
Text Types	Texts include the following: - evaluative texts such as editorials, commentaries, criticisms, opinion pieces, political analyses, reviews, apologia; - expository texts such as (a) essays to explain a problem, (b) discussions of notions, ideas or concepts, or (c) specialized texts that discuss

	concrete but technical issues intended for a non-specialist reader; - argumentative texts that support, defend or refute a policy, strategy, or program such as a proposal by a political party, articles to argue in favor of or against an issue, etc.
Text Organization / Text Characteristics	Texts contain analysis, value judgments or evaluation of things and events against a backdrop of shared information/knowledge. The language contained is used to accommodate author's message, i.e., through writing styles/personalized expressions, use of given inferences, hypothesis to invite readers to evaluate the material. Facts are usually selected and serve to convey a point of view. These phenomena are sometimes referred to as "shaping." The author of the text provides background information necessary for readers to understand the argumentation; the author assumes that the reader shares some basic knowledge of the topic to be discussed and the circumstances leading to its discussion. Texts contain familiar sociolinguistic and target-language cultural references, although this content is not a primary focus in conveying the text's meaning. Texts also display a wide range of discourse structures and show clear cohesive discourse with cohesive devices characteristic of the target language.
Lexical Range	Lexical Range is sufficient to allow for detailed analysis of, opinions about, and commentary on abstract, societal-related topics and for discussions of concrete, technical issues. Lexical Range tends to be topic-specific and covers the concrete and abstract domains. Use of idiomatic expressions to express the author's message is becoming more evident.
Syntactic Complexity	Sentence structures become more complex, that is, the (extensive) use of complex and compound sentences (in order to express ideas, concepts or notions about abstract topics). All tenses and aspects, moods e.g., conditional and subjunctive), and voices (e.g., active and passive) are represented. Texts display the full range of sentence structures, i.e., compound, complex, and compound-complex forms. Texts may also present the full range of verbal forms, e.g., tense, aspect, mood, voice, and of nominal (case) forms.
Length	Up to 400 words according to the English rendering
Sources	The Internet and print media including newspapers, magazines/journals, books, essays for the general public
Focus of Task/Question	Understanding - Main ideas - Major details - Lines of argumentation supporting the author's view - Author-intended implications / inferences - Vocabulary/idiomatic expressions in context - Author's attitude / position / point of view / tone

	Comparing points of view as presented in the text Drawing appropriate conclusions
Cognitive Load	It is expected that examinees may experience "some difficulty with unusually complex structure and low frequency idioms, they are able read "within a normal range of speed," and "rarely have to pause over or reread general vocabulary." They are expected to demonstrate well-developed language knowledge of the target language including the socio-cultural aspects. Examinees are required to process abstract information on socio-political issues. They are required to understand, analyze, and integrate the information presented and identify the correct option in MC items or provide the correct answers in CRT items. The relationship between the target language input and its associated questions and expected responses is less transparent as expected responses tend to require examinees to synthesize information, and express their judgments and opinions about the target texts in ways that bear little direct relationship to the texts those judgments and opinions are based on.

ILR Level	**Reading - Level 36**
ILR Skill Level Descriptions	Can comprehend a variety of styles and forms pertinent to professional needs. Rarely misinterprets such texts or rarely experiences difficulty relating ideas or making inferences. Able to comprehend many sociolinguistic and cultural references. However, may miss some nuances and subtleties. Able to comprehend a considerable range of intentionally complex structures, low frequency idioms, and uncommon connotative intentions, however, accuracy is not complete. The individual is typically able to read with facility, understand, and appreciate contemporary expository, technical or literary texts which do not rely heavily on slang and unusual items. [Data Code 36]
Description of Ability	Able to - read with facility, understand, and appreciate contemporary expository, technical, or literary texts that represent a variety of styles and forms; rarely misinterprets such texts. - comprehend many sociolinguistic and cultural references. - comprehend a considerable range of intentionally complex structures, low frequency idioms, and uncommon connotative intentions.
Skills to Be Assessed	Ability to - understand main ideas / details - understand the author's arguments in support of his/her position - understand implications or inferences intended by the author - understand different points of views presented in the text - understand vocabulary/idiomatic expressions in context - draw appropriate conclusions / summarize - identify the writer's tone / attitude / position
Target Language Input	
Text Mode / Purpose	Mixed Evaluative/Projective Mode (Child) – Texts are written with a strong social / public purpose similar to those at Level 3, but some idiosyncratic approaches to the handling of the subject matters begin to appear.
Text Types	Text types are similar to those at Level 30, i.e., evaluative texts, expository texts and argumentative texts. Literary texts may be included depending on the convention of the target language culture. However, such "literary texts" are generally related to writing or literature, e.g., commentary or analysis. "Literature" per se, i.e., poems, short stories, etc. written by authors to be read as their own works, is generally not included on DLPT5s.
Organization / Text characteristics	Texts contain analysis, evaluation, or hypothesis of complex subject matters / problems / topics. The author assumes some shared background knowledge from the reader. But texts can be culturally dense.

	Understanding the author's message is increasingly linked to culture-specific content / allusions in the text. Texts show a sophisticated use of cohesive devices and display a variety of text structures to convey a greater variety of discourse functions. Some writings may display literary creativity / idiosyncratic approaches to the treatment of the subject matter / topic. Author may use some less familiar cultural references or dense sociolinguistic information to express his/her points of view. Author may use a considerable range of intentionally complex structures, low frequency idioms, and uncommon connotative associations.
Lexical Range	Texts present a wide range of Lexical Range to allow for detailed analysis / evaluation of the topic under discussion. To achieve the author's communicative intent, texts may make use of commonly and/or less frequently used idiomatic expressions and/or slang. However, texts do not rely heavily on slang or infrequently used idioms for that purpose.
Syntactic Complexity	Texts contain the full range of syntactic structures. The use of sentences to express particular pragmatic functions or register becomes evident. Structural features may be specifically chosen to express particular rhetorical functions or meanings.
Length	Up to 400 words according to the English rendering
Sources	The Internet and print media including newspapers, magazines/journals, books, essays for the general public
Focus of Task / Question	Understanding - main ideas / major details - points in support of author's lines of argumentation - various viewpoints for or against author's arguments - inferences or author-intended implications - vocabulary / idiomatic expression in context Drawing appropriate conclusions or summarize Author's attitude / tone / position on an issue
Cognitive Load	Though examinees may "miss some nuances and subtleties" and their "accuracy is not complete," they are able to "comprehend a variety to styles and forms pertinent to professional needs" and "comprehend many sociolinguistic and cultural references." At this level, examinees are expected to demonstrate well-developed language knowledge of the target language including cultural knowledge. Examinees are required to process the information in the target texts and also understand the issue/topic being discussed and the author's intentions from the target-language culture's point of view. Examinees are required to synthesize the ideas presented by the author, and identify the correct option in MC items or provide the correct answers in CRT items. The scope of relationship between the target language input and its associated questions and expected responses is wide as more expected responses require examinees to express judgments and opinions about the

	target texts in ways that bear little direct and one-to-one relationship to the texts those judgments and opinions are based on.

ILR Level	Reading - Level 40
ILR Skill Level Descriptions	Able to read fluently and accurately all styles and forms of the language pertinent to professional needs. The individual's experience with the written language is extensive enough that he/she is able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references. Able to "read beyond the lines" (that is, to understand the full ramifications of texts as they are situated in the wider cultural, political, or social environment). Able to read and understand the intent of writers' use of nuance and subtlety. The individual can discern relationships among sophisticated written materials in the context of broad experience. Can follow unpredictable turns of thought readily in, for example, editorial, conjectural, and literary texts in any subject matter area directed to the general reader. Can read essentially all materials in his/her special field, including official and professional documents and correspondence. Recognizes all professionally relevant vocabulary known to the educated non-professional native, although may have some difficulty with slang. Can read reasonably legible handwriting without difficulty. Accuracy is often nearly that of a well-educated native reader. [Data Code 40]
Description of Ability	Able to - read fluently and accurately all styles and forms of the language pertinent to professional needs - relate inferences in the text to real-world knowledge and understand almost all socio-linguistic and cultural references - read and understand the intent of writers' employment of nuance and subtlety. - "read beyond the lines," that is, understand the full ramifications of texts as they are situated in the wider cultural, political, or social environment. - follow unpredictable turns of thought readily, for example, in editorials, conjectural, and literary texts in any subject matter area directed to the general reader.
Skills to Be Assessed	Ability to - understand the full ramifications of the text as it is situated in a wider context, i.e., read "beyond the lines," and draw appropriate conclusions / inferences - understand major points supporting the author's argumentations - understand subtle author-intended implications - understand the nuances / subtleties employed by the author - understand the significance of the socio-linguistic or cultural references in the text - understand vocabulary / idiomatic expressions in context and/or the significance of the author's choice of such vocabulary / expressions - follow turns of thought in the text that are unexpected

	- detect the author's attitude / tone / points of view
Target Language Input	
Text Mode / Purpose	Projective Mode (Child) – Texts are written for social, public or personal purposes. Texts contain Instructive or Evaluative elements; however, the particular approach in the handling of the topic suggests the texts or authors (1) take a novel or creative approach to (the examination / treatment of) a problem / topic, or (2) offer (highly) individualized insights to issues regarding the human race or of interest to the author, or (3) display idiosyncratic language use or cultural values that lie outside those that are widely familiar.
Text Types	Texts include think-pieces, philosophical expositions, satiric writings, writings with sophisticated humor, or individualized writings of a colloquial nature, etc.
Organization / Text characteristics	- Texts display dense cultural and linguistic information. The authors assume that the reader brings a great store of background knowledge to the reading task. Therefore, they leave historical, cultural or other references unexplained in the texts. Texts are highly individualized, idiosyncratic, original, and/or culture-bound and demand a great deal of reader input. - Texts contain highly individualized or culture-specific forms of discourse, abstract metaphors, and symbolism. - Texts also demonstrate the author's virtuosity with language and may mix uses of formal and informal registers to achieve subtlety and nuances. - Texts display the author's unique way of thinking, clearly show the author's tone, are cogently persuasive, but challenge the reader due to the author's unpredictable turns of thought / unexpected means of argumentation. - Texts display a wide range of contexts from colloquial to careful and formal.
Lexical Range	Texts exhibit a full range of vocabulary – from concrete to abstract, and from colloquial to formal – as it would be known to a well-read native speaker of the language. Vocabulary / idiomatic expressions may be used in a novel way to achieve the author's communicative intent, and the choice of given vocabulary and/or idiomatic expressions is so appropriate that attempts at substitution / replacement are not possible. The author may also choose slang or less commonly used idioms as a means to convey his/her communicative intent.
Structural / Syntactic Complexity	The author employs a full range of syntactic structures and uses them proficiently to express complex ideas. Texts at this level generally display idiosyncratic choices in the use of syntactic features.

Length	Up to 500 words according to the English rendering
Sources	The Internet and print media such as periodicals/journals, books, collections of writings, essays, monographs, etc. Such sources present content that caters to well-read target language users.
Focus of Task / Question	- Understanding subtle, author-intended implications / inferences - Understanding major points supporting the author's argumentation - Understanding the author's attitude / tone - Understanding vocabulary / idiomatic expressions / significance of the socio-linguistic and cultural references in context - Drawing appropriate conclusions / synthesizing
Cognitive Load	The extent of language knowledge expected from the examinees is almost equivalent to that of the native speakers of the target language. Examinees are required to fully understand the content and context of the text, to synthesize the information and ideas, and identify the correct option in MC items or provide the correct answers in CRT items. The scope of relationship between the target language input and questions with their expected responses is wide as expected responses requires examinees to express their judgments and opinions on the basis of their knowledge of the target language culture about the target texts in ways that have little direct relationship to the target texts those judgments and opinions are based on.

As regards the ILR Skill Level Descriptions for Reading, there are a number of statements that describe abilities which are not operationalized in the DLPT5. In the following paragraphs, we identify these statements and offer explanations as to why the abilities they describe are not tested.

ILR Level 1+ contains the statement, "Can guess at unfamiliar vocabulary if highly contextualized, but with difficulty in unfamiliar contexts." DLPT 5 items do not test discrete vocabulary at this or at any other level. We are not testing the ability to "guess" in terms of Lexical Range; this statement rather is a reading strategy. Hence, this ability is not measured in the DLPT5.

Regarding the statement in the Level 2 descriptor, "… persons who have professional knowledge of a subject may be able to summarize or perform sorting or locating tasks with written texts that are well beyond their general proficiency level" the DLPT5 test is a test of general proficiency. Examinees' performance related to specialized knowledge of a subject matter is not the focus of assessment. The content of DLPT5 texts is kept at a level for the general readers.

The ILR descriptor for Level 2+ makes reference to "… texts which could only be read by persons with much higher proficiencies." This specific reference speaks to texts that are beyond level 2+ but is generalizable to all ILR levels tested on DLPT5, i.e., texts with characteristics of any level higher than the one being tested are not included.

ILR Level 2+ also makes reference to "topics related to special professional interests" and contains the statement, "Is markedly more proficient at reading materials on a familiar topic." Texts on DLPT5 are not selected specifically for examinees or presented to them because of an examinee's professional background or experiences.

In the ILR Level 3 statement, "Can get the gist of more sophisticated text, but may be unable to detect or understand subtlety and nuance," the references to "more sophisticated texts" and "subtlety and nuance" speak to texts that are beyond Level 30. Such texts would not appear at this level as noted above. However, this ability to understand language at a higher level is accounted for in the calibration of the examinees' final scores.

Regarding the ILR Level 4 statement, "Can read essentially all materials in his/her special field, including … correspondence," and the qualifier "pertinent to professional needs" in the statement, "Able to read fluently and accurately all styles and forms of the language pertinent to professional needs," texts on DLPT5 are not selected specifically for examinees or presented to them because of an examinee's professional background or experiences.

The statement, "Can discern relationships among sophisticated written materials on the context of broad experience," is, in one sense, subsumed under the concept of "reading beyond the lines" as mentioned in the Description of Ability section for ILR Level 4. However, this statement should not be understood to mean that examinees are asked in DLPT5 to read several topically related texts simultaneously and to demonstrate understanding of the interrelationship between / among such texts. In DLPT5 examinees read and answer questions about only one text at a time.

Regarding the ILR Level 4 statement, "Can read reasonably legible handwriting without difficulty," DLPT5 does not include any hand-written texts. Such reading would be considered a measure of performance, not of proficiency.

3.3 Listening Comprehension

In this section, the operational definition of listening comprehension as measured in the DLPT5 will be provided. This section includes a general discussion of listening ability, of the factors affecting listening comprehension, and of the ways in which these factors are dealt with in the test design. These factors include examinee characteristics, characteristics of the spoken texts, standard languages and dialects, accents, authenticity, and question characteristics. Lastly, we will present the interplay of the different factors with ILR level and describe the characteristics of the listening texts and question types in the DLPT5 Listening test.

3.3.1 What is listening?

Listening comprehension is an important and complex cognitive skill. It is also perceived as a difficult skill to attain by language learners. Currently, there is a lack of an agreed-upon definition of what listening comprehension is in either the first-language or second-language context. A general consensus among testing researchers with regard to listening comprehension is that it involves an interaction between linguistic codes and the examinees' ability to process the acoustic input and construct some type of mental representation on the basis of their linguistic knowledge, world knowledge, and personal experience (Lund, 1991; Buck, 2001). Listeners process what they hear in real time and simultaneously continue taking in new acoustic input that needs to be structured. Listeners do both at a pace set by the speaker(s), over which they often have little or no control. Three types of knowledge are activated to help listeners process the incoming acoustic signal: declarative knowledge, procedural knowledge, and background knowledge. Declarative knowledge is what the listener knows about the target language, its structures, and its functions. The listener has conscious control of his or her declarative knowledge and applies it when using the target language. Procedural knowledge, on the other hand, is the ability to use the target language effectively in an automatic manner. Background knowledge orients the listener in a given listening context with information stored in his or her long-term memory. It also helps the listener interpret what he or she hears.

Listeners then use the types of knowledge described above to transform those signals into sets of mental propositions. Because of individual differences in knowledge, working memory capacity, and cognitive processing, listeners exposed to the same acoustic signals may arrive at different sets of propositions. On a listening comprehension test, examinees then further process this set of mental propositions to produce an appropriate response to a question based on aural input. On the constructed-response DLPT, this response is written (involving the examinee's writing ability); on the multiple-choice DLPT, this response is a selection from among a set of choices (involving the examinee's reading ability). Apart from the listener's knowledge and cognitive abilities, how well a response demonstrates comprehension is also partly dependent on how well the listening text is presented or represented. In short, listening comprehension results from the interplay of the listener, the listening texts, and the questions.

3.3.2 Defining listening comprehension

We regard listening comprehension in a testing situation as the product of the interplay of factors such as listeners' language ability, text difficulty, and question difficulty. In this section, we will define (1) the scope of listener ability measured in the DLPT5 Listening Test; (2) the types of listening texts that exemplify features of real-world spoken texts; and (3) the types of questions

that engage examinees in the various sub-skills essential for successful listening. As discussed in section 3.1, the ILR Skill Level Descriptions provide a general account of the range of listening ability and its sub-skills required of second-language users in different listening situations. In their selection of listening texts, DLPT5 test developers follow guidelines in the textual descriptions of the ILR Skill Level Descriptions, supplemented by Child's text typology, inferences from the skill level descriptions about successful functioning, and functional inferences from the skill level descriptions for other skills. The ILR Listening Skill Level Descriptions also set the standards for what second-language listeners can do at each of the proficiency levels, and so they are also used as guidelines in identifying the sub-skills to be tested at each level and in developing the tasks to test these sub-skills, again with supplementation from the methods mentioned above.

For example, the Listening Level 1 description states that the listener is able to "understand utterances about basic survival needs and minimum courtesy and travel requirements in areas of immediate need or on very familiar topics," and "understand simple questions and answers, simple statements, and very simple face-to-face conversations in a standard dialect." The Listening Level 1 description also suggests that listeners understand "main ideas," and their knowledge of the target language allows them to understand variation of sentences with "[a] similar level [of] vocabulary and grammar." These are general statements about how well a Level 1 target language user functions in the target language environment and what kinds of texts he or she can deal with. We can compare these ability descriptions with those in the reading and speaking skill level descriptions, since the functional contexts in which second-language users must produce or comprehend texts at a given level are the same. At Reading Level 1, the ILR skill level descriptions state that the reader "can read … simple language containing only the highest frequency structural patterns and vocabulary and "can get some main ideas and locate prominent items of professional significance in more complex texts," and "understand known language elements that have been recombined in new ways to achieve different meanings at a similar level of simplicity." The Level 1 description for speaking says, "the individual can typically satisfy predictable, simple, personal and accommodation needs." These descriptors suggest, adding inferences from Child's modes and functional considerations, that Level 1 listeners are generally capable of understanding main ideas and identifying specific information important for immediate needs. Hence these two sub-skills are listed as Level 1 requirement, and all questions at level 1 focus on these two sub-skills. In addition, the type of texts to be selected, collections of mostly simple sentences with basic survival vocabulary and little internal organization such as would be commonly encountered in survival situations, is inferable in this way.

Another influence on task development, aside from the sub-skills, is the accuracy statements found in the ILR Skill Level Descriptions. Misunderstanding or failure to accomplish certain tasks happens as a result of examinees' inadequate knowledge in the target language. For example, at Level 1, misunderstanding due to "overlooked or misunderstood syntax and other grammatical clues" is evident. At Level 2+, listeners "may display weakness or deficiency due to inadequate vocabulary base or less than secure knowledge of grammar and syntax." The fact that examinees may misunderstand or fail to understand the spoken texts throughout the ILR levels is accounted for by the amount and precision of information required to complete the tasks successfully.

One major problem presents itself with regard to listening comprehension as compared to reading comprehension: the issue of participatory versus non-participatory listening. For example, the ILR descriptions at Level 1 state that the listener's understanding is based on a sympathetic participant who speaks "clearly" and "at a rate slower than normal with frequent repetitions or paraphrase." At Listening Level 2, the listener is "able to understand face-to-face speech in a standard dialect, delivered at a normal rate with some repetition and rewording, by a native speaker not used to dealing with foreigners, about everyday topics, common personal and family news…" These ability requirements suggest the ILR addresses listener ability at Levels 1 and 2 in the context of participatory listening, in which the listener and speaker are able to negotiate meaning. The listener is able to ask for clarification and the speaker is able to accommodate the listener through means such as slowed speech, repetitions, rewording, or syntactic simplification. The ILR Skill Level Descriptions address listener ability in non-participatory listening only starting from Level 3, at which the listener "can follow accurately the essentials of … radio broadcasts, news stories…" The design of the DLPT5, however, for practical reasons, only tests non-participatory listening. The disconnect between participatory listening as described in the ILR Skill Level Descriptions and non-participatory listening measured in the DLPT5 Listening Test thus necessitates a high degree of supplementation using the techniques outlined above in order to specify task and text selection parameters.

Table 3-3 presents the listening framework for the design of DLPT5 Listening Test. Detailed descriptions of listening abilities and text and task characteristics are provided in Table 3-4.

Table 3-3

VI.	**Purpose**: to measure US civilian and military language analysts' listening comprehension ability in non-participatory listening in a foreign language
VII.	**Construct**: the extent of the ability as specified by the ILR to • process authentic spoken language automatically and in real time, • understand the linguistic information that is unequivocally included in the spoken text and • make any speaker-intended inferences that are unambiguously implied by the content of the text or by intonation, accent, the use of the tone, or other oral characteristics.
VIII.	**Characteristics of text and question**: presented in Table 3-4
IX.	**Characteristics of examinees**: See Section 2.5
X.	**Listening abilities**: presented in Table 3-4; listening abilities as characterized according to the ILR Skill Levels Descriptions (Appendix A).

- Identifying the examinee, text, and question characteristics

The examinee
Examinees' proficiency level, background knowledge, memory, and individual attributes, as well as strategies they use during tests can have considerable impact on their listening comprehension ability and test performance.

Learners at lower levels of proficiency tend to apply their linguistic knowledge to understand the listening texts (Teng, 1999) whereas higher-ability learners use other strategies to monitor their comprehension (Goh, 2002). Also, learners use linguistic knowledge to decode texts where there is little background information available.

Studies show that different listening strategies are used by people of different proficiency levels. Goh (2002) studied listening strategies and identified some cognitive strategies including inferencing, elaboration, prediction, translation, contextualization and visualization, and metacognitive strategies such as self-monitoring, comprehension monitoring, selective attention and self-evaluation. Her findings show that while both the higher ability and lower ability groups of listeners use similar strategies, the higher ability listeners make more effective use of the cognitive and metacognitive strategies. Vandergrift (in Rubin, 1994) also found that more proficient listeners make greater use of metacognitive strategies, whereas less proficient listeners rely more on cognitive strategies.

Research findings in second-language listening indicate a delicate interaction between top-down (or higher-level) and bottom-up (or lower-level) processing. Bottom-up processes refer to processes like perception, word recognition and sentence/utterance parsing. Listeners use bottom-up processes when they construct meaning from the phoneme-level to discourse-level features. Top-down processes refer to the use of cognitive or metacognitive skills to build a conceptual framework for comprehension based on context, background knowledge. For competent listeners, the bottom-up processes are automatic and the top-down processes are more controlled, i.e., conscious. Low-level comprehension processes such as word recognition and syntactic parsing can become automatic and will free more cognitive capacity for higher-level processing such as making inferences. When there is difficulty processing at word level, there will be little cognitive capacity left for processing the meaning or intent of a given text. This notion is supported by Conrad's study (1985), which indicated that as learners' proficiency decreases, they tend to rely more on syntactic cues than on contextual and semantic cues. These findings are reflected in the types of texts and tasks described in the ILR Skill Level Descriptions and found on the DLPT5. As the ILR level increases, an increasing amount of higher-level processing is required in order to reach a level-appropriate understanding of the texts and complete the level-appropriate tasks successfully.

Background or real-world knowledge is also an important facilitator in listening comprehension. Empirical studies exploring the relationship between prior knowledge and listening comprehension have shown that background knowledge improves listening comprehension (Chiang & Dunkel, 1992; Long, 1990; Schmidt-Rinehart, 1994). Examinees are expected to use their real-world knowledge and their knowledge of the target-language culture to help them understand the listening passages. However, since we cannot assume any particular subject-matter expertise on the part of examinees, passages that require subject-matter expertise for comprehension are not selected for the test. In order to control for the effect of subject matter knowledge, DLPT5 test developers select passages originally intended for the general audience, and they include as wide a variety of topics as possible. The content of the DLPT5 listening passages is classified according to the FLO content areas (Table 2-1), and the distribution of content on the DLPT5 is determined with reference to these classifications.

The listener's memory capacity may also influence the listening process. Incoming aural stimuli are stored in the listener's working/short-term memory. New information then is integrated with pre-existing information that is stored in the listener's long term memory. Researchers in cognitive psychology suggest that information is held in the short-term/working memory for up to 30 seconds, after which time, the information is lost if it is not reinforced. Rost (1990) wrote that 30 – 60 seconds are required by working memory to sort out information in the aural stimuli. Information that activates knowledge stored in the listener's long-term memory gets processed rapidly. If the aural stimuli consist of predominantly new information, more time and space is needed for processing this information, which leads to limited access to old information. Short-term memory has been claimed to play a role in information-processing tasks. However, research findings have indicated mixed results. Dunkel et al. (1989) find that people who have good short-term memory recognize significantly more concept information and detailed information than people who have poor short-term memory. The study done by Carrel et al. (2002) on the effects of note-taking and lecture length in the TOEFL 2000 listening test suggested that short-term memory had an insignificant effect on performance on the computer-based listening test. Current understanding suggests that listening efficiency is largely based on the level of the listener's language proficiency. A competent listener processes and retrieves information as well as achieves understanding efficiently and effortlessly. Yet, in order for examinees to better process information presented in the test passages, DLPT5 test developers avoid developing tasks that target information that could be considered insignificant in light of the purpose and content of the listening passage, since insignificant information tends to be less memorable than significant information. The role of short-term memory on DLPT5 test performance may need further investigation.

Lastly, the effect of note-taking on listening test performance has been considered. Study findings on note-taking are inconclusive. Some suggest that note-taking has some benefits over recall of information for individuals with better short-term memory, aids performance of certain task types such as problem-solving and application, and helps when listening texts are longer than 5 minutes. Other studies suggest that there is no significant relationship between note-taking and no note-taking (Cheuvront, 2004). Some debilitating effects of note-taking were uncovered; one of them was that students are only capable of recording 20 words per minute, whereas most lectures are spoken at a much faster rate. When the rate of speech reaches 135 words per minute or greater, students who try to take notes perform worse than students who do not take notes (Ladas, in Cheuvront, 2004). It was decided that note-taking should not be allowed for the lower-range MC tests because the answer choices provide clues, but that note-taking would be allowed for the CRT tests for two reasons: first, for the practical reason that examinees are allowed to begin typing their answers as soon as the passage begins playing; second, because the task of producing information seems to be a greater burden on memory than the task of recognizing information. Further investigation into note-taking in the DLPT tests is necessary.

The listening texts

Characteristics of listening texts that can affect test performance include the lexical and syntactic aspects of the aural input; the speech segment's acoustic variables, such as accent, speech rate, pauses, phonological modification, and stress and rhythmic patterns; the text type, for example, dialogues, news reports, lectures, etc.; and the degree of authenticity as perceived by the examinee.

Buck (2001) noted that spoken texts, unlike written texts, consist of short and clause-like idea units about 7 words long and 2 seconds in duration, generally strung together by coordinating conjunctions and relatively simple in their syntax. He also pointed out that spoken texts include a variety of linguistic and prosodic features such as use of dialect, slang and colloquialisms, accent, fillers, false starts, hesitation, self-correction etc. Therefore, a listening comprehension test should include features of authentic spoken texts as well as a range of text types on the oral continuum to reflect listening in the real world. DLPT5 test developers address this issue in two ways. First, they select, whenever possible, texts from authentic, real-world sources, which tend to exhibit these features of spoken language. Second, they avoid any texts that, although presented in the aural mode, were originally produced with the intention that people would read them rather than hear them (e.g., a newspaper opinion essay read aloud).

Vocabulary load affects difficulty in listening comprehension (Dunkel, 1991). Thompson (1995, in Bejar et al, 2000) proposed that texts containing predominantly high-frequency vocabulary will be easier to understand than those which include jargon and technical words. Colloquial words and phrases may be the most difficult to comprehend due to phonemic reduction – the fast, slurred pronunciation of those words by native speakers make them difficult to understand (Ur, 1984). The presence of abstract words and vague words such as 'very' or 'some' in the text may also influence task difficulty (Power, 1985). In the DLPT5 context, this aspect of texts is generally controlled by matching the range of vocabulary to the levels and types of spoken texts as specified by the ILR and text typology. For example, at Level 1, the areas of comprehension only include "meals, lodging, transportation, time and simple directions," which refer to basic needs. Vocabulary used in texts dealing with language use in these contexts includes basic, high-frequency words. At Level 2, the listeners' ability to comprehend spoken texts expands to include "everyday topics, common personal and family news, well-known current event, and routine office matters through descriptions and narration about current, past and future events." Vocabulary used in texts dealing with language use in these situations goes beyond words for basic needs and includes a wider variety of words of a concrete nature. There may be some use of high-frequency idiomatic expressions. But the intended meanings of words used are straightforward. At Level 3, the contexts for language use for the listeners is even wider, so Level 3 texts tend to include a wide spectrum of vocabulary associated with different text types. Refer to Table 3-4 for the range of vocabulary in the DLPT5 listening passages.

Speakers often use pauses and changes of speed to provide clues for the chunking of information (Rubin 1994). Pauses and fillers such as 'um' are thought to facilitate comprehension (Bygate, 1987); however, research results are not conclusive. These factors are not controlled for on the DLPT5, although native speakers of the target language must judge the listening texts to sound natural.

Speech rate is an important variable affecting listening. Tauroza and Allison (1990, in Buck 2001) looked at the average speech rates for British speakers in radio monologues, conversations, and interviews aimed at native speakers of English. They found the rates varied among text types with conversation being the fastest at 210 words per minute, interviews at 190 wpm and radio monologues at 160 wpm. Flowerdew (1994, in Bejar et al, 2000) reported that slowing down speech to 100 – 150 wpm does not aid in comprehension. But when speakers speed up their

speech, the result is reduced comprehension. This finding parallels Griffiths' earlier study on speech rate (Griffiths, 1990). These studies used ESL/EFL learners; Griffiths suggested that different languages may have different 'normal' speech rates, and, of course, normal rates vary among text types and individuals. Given these results, we would expect differences in difficulty among texts at the same ILR level that are spoken at different rates. Note that in order to accommodate statements from the ILR Skill Level Descriptions, some DLPT5 passages at lower ILR levels may be spoken at a slower-than-normal rate of speech, but they must still be judged by a native speaker as sounding natural. Unnaturally slowed, or "teacher-talk," passages are avoided.

Another factor that influences passage selection is the acoustic quality of the sound file. Background noise and fidelity of the signal are two factors, for example that can profoundly affect the ability of examinees at various levels to extract the meaningful elements from the aural signal. The ILR Skill Level Descriptions explicitly refer to the quality of the aural signal at several levels. For example, at Level 1, the statements the listener can understand "must often be delivered more clearly than normal". At Level 2, the listener "only understands occasional words and phrases of statements made in unfavorable conditions, and at Level 4, the listener still "has difficulty in understanding…speech in unfavorable conditions." The factor of background noise is not explicitly mentioned in the ILR Skill Level Descriptions, and research on the effect of this factor on listening comprehension has been inconclusive. In order to address these factors, DLPT5 test developers adhere to the standard that the words in the passage need to be generally clearly audible and identifiable for a non-native listener at the targeted level, and that any parts that are not clearly audible cannot be the basis for any of the tasks based on that passage.

The degree of authenticity of the spoken text is another important aspect affecting examinee performance. Authentic spoken texts include variation in accent, stress and intonation, and phonological modification of lexical items. DLPT5 test developers select passages with different acoustic variables according to the context of the aural input. Research findings have supported the hypothesis that unfamiliar accent causes listening difficulty. The ILR Skill Level Descriptions explicitly mention the factor of regional variation at various levels. For example, at Levels 1, 2, and 3 listeners can understand text types appropriate to the level "in a standard dialect." In addition, at Level 3, the listener "does not understand native speakers if they…use some slang or dialect." At Level 3+, the listener has an "increased ability to understand native speakers…using nonstandard dialect or slang; however, comprehension is not complete." Finally, at Level 4, the listener can understand all types of speech "in all standard dialects" and "the essentials of speech in some non-standard dialects," but "has difficulty in understanding extreme dialect and slang." On the DLPT5, dialects and accents considered to be standard are used throughout the ability levels, except at the highest levels, in which listening ability includes the ability to comprehend texts delivered in non-standard versions of the dialects and accents. The perception of what the standard dialects and accents are may differ in speech communities in different geographical regions. In cases where the target language is spoken in different geographical regions, any accent considered to be standard can be used. DLPT5 test developers control for regional variation in the use of grammatical and phonological features by including passages that all target language reviewers agree sound natural. In addition, listening passages also include participants of both sexes and of different age groups and professions to reflect authentic language use in the real world.

In short, regarding authenticity of the spoken text, it is important to select texts for use on the DLPT5 that exhibit features of typical, current spoken texts that can be found in the target-language use situation. For this reason, wherever possible, fully authentic texts produced by users of the target language intended to be heard by other users of the target language in the target-language culture are selected. There are, however, various circumstances in which texts must be used that are not fully authentic.

For example, the fully authentic text might significantly exceed length constraints for test passages at a given level, or it might contain embedded off-topic material. In these cases the fully authentic texts might be abridged for use as test passages, as long as the test passage remains natural and coherent. Also, depending on the target language, material in certain content areas might be difficult or impossible to collect. The same might also be true for certain text types at certain levels. Additionally, certain samples might be appropriate linguistically but have poor sound quality.

Regarding passage length, every effort is made to adhere to the limits as shown in the DLPT5 Test Specifications. However, in some cases, slightly longer passages may be allowed for use in an operational test if they are needed to meet other requirements, for example appropriate topical coverage. In such cases, the item calibration information must confirm that those passages and their related test items functioned as expected and were deemed of sufficiently high quality for inclusion in an operational test form.

In order to achieve a desirable sampling of text types and subject areas at each level or acceptable sound quality, it is sometimes necessary to purpose-produce listening texts in the studio. There is a continuum from fully scripted to fully unscripted speech that can be produced in a studio setting. For an example of fully scripted speech, target-language experts might script a dialogue or interview on a certain subject and then read the script as naturally as possible in the studio recording session. For an example of fully unscripted speech, target-language experts might be given a topic in the studio recording session and asked to discuss it on the spot. If the sound quality of an original authentic text had been poor, target-language experts might be asked to revoice a transcript of the original in the studio recording session. Or target-language experts might be asked to familiarize themselves with certain content or make an outline for themselves and then have a discussion based on that knowledge or those notes. In these studio recording sessions, certain texts might also be rerecorded in order to enhance or eliminate certain linguistic features.

In the end, we rely on the expertise of target-language users who have extensive knowledge of the target language and culture and can judge whether a given test passage could appear in the appropriate context for its text type and be regarded as natural by other members of the target-language use community.

Shohamy and Inbar (1991) considered comprehensibility of three text types: a news broadcast, a mini-lecture and a consultative dialogue. They found that the news was the most difficult, followed by the mini-lecture, with the dialogue being the least difficult text type. The study by Brown et al. (1985, in Rubin 1994) on first-language English-speaking students suggested that

narrative texts are easier to listen to and recall than expository texts. In addition, their data suggested that "events described in chronological order are easier to recall than narratives with disrupted sequences or flashbacks." These findings are reflected in the text types found at the different ILR levels.

Although the DLPT5 Listening Test measures non-participatory listening comprehension ability, the test content represents realistic listening in terms of contexts and text/discourse types. The DLPT5 Listening Test addresses the issue of context by including listening passages in a variety of social and business situations. Three discourse types have been identified: planned speech, semi-planned speech, and unplanned speech. Planned discourse is prepared, organized or polished listening texts of the sort usually found in news reports; unplanned discourse is spontaneous speech produced without taking much time for preparation or organization. In the middle is the semi-planned discourse in which the speaker has some time to prepare a mental draft or to practice beforehand, such as giving a presentation or some kinds of interviews. These types have different linguistic features that vary in syntactic complexity. Some types are easier in terms of comprehensibility than others. But these text types are part of the target language use and are all represented in the DLPT5 test, as far as possible. Refer to Table 3-4 for a description of listening text types included in the test.

Listening is a purposeful process (Rost, 1990). When people listen for a purpose, this purpose will drive the comprehension process. Brown and Yule (1983) identified two listening purposes along the listening continuum: interactional and transactional. Listening to interactional discourse means the listener is being an active participant in collaborative discourse and the purpose is to communicate "good will." In such types of listening, the listener's ability to display understanding and signs of participation in expected ways is important. Transactional purposes of listening refer to instances of listening in which the listener and speaker are engaged in information transference. The listener is usually expected to take note of the information, such as writing down directions to a friend's house or taking lecture notes. The listener can clarify the information if he or she wishes to. In cases in which the listener is not able to interact with the speaker, the speaker generally goes to considerable length to make sure his or her message is clear. However, in this type of transactional listening (or non-participatory listening), the listener is unable to use clarification strategies directly, and his or her understanding is not normally explicitly displayed (Rost, 1990). The DLPT5 Listening Test engages examinees in the transactional type of listening in which examinees are required to listen for specific information, listen for basic comprehension, i.e., main ideas, and listen to integrate and connect detailed information to make a coherent whole and understand the intent of the speaker(s). Interactional listening is represented by non-participatory listening to interactional dialogues.

Henning (1991) studied passage length and memory load on TOEFL listening passages. Findings suggested that length did not contribute to difficulty, but repetition of the passage tended to make the tasks easier. He also suggested test reliability increases with longer passages, and there is no evidence that any additional burden on memory associated with passage length would negatively affect performance on the tasks. The length of passages in the DLPT5 naturally becomes progressively longer as the levels of passages go up, based on the nature and content of the texts as observed in the target language use domain. However, passage length must be limited by practical considerations of test length; therefore, maximum lengths for passages at each level

have been established (see Table 3-4). Within a given ILR level, the passages can vary somewhat in length. DLIFLC is in the process of developing research proposals to establish the relationship of passage length and retention in both native and second languages and apply the findings to DLPT5 listening passages.

The number of times the aural input is played has been discussed between the DLPT5 test developers and test users. It was decided for practical reasons of overall test length that for the MC test, listening passages are played once at Levels 1 and 1+ but twice for Level 2 and above, and for CRT tests, each passage is played twice. It is believed that the second playing may help examinees process the content more efficiently. Passages at Levels 1 and 1+ are short, and the content of passages generally relates to everyday encounters with repetition of information built in. In addition, the examinees are able to see the answer choices, which, to some extent, provide clues to the content of the passages. The reason for CRT passages to be played twice is that examinees are required to produce written answers. The second playing of the passages is to help examinees process the information and formulate their answers even from Level 1. Investigation on the relationship between the number of playings and listening difficulty is necessary for future generations of the DLPT tests.

The DLPT5 test developers examine the spoken text from the perspectives of discourse, linguistic and paralinguistic features. Discourse features concern the pragmatic aspect of language use, i.e., the purpose, organization, types and length of the spoken text. Linguistic features relate to vocabulary and syntactic complexity of the spoken text, and paralinguistic features describe the phonological and acoustic aspects of the spoken text like sound quality, speech rate, dialects and accents as well as oral characteristics found in spoken texts such as the use of fillers, pauses, repairs, omissions, assimilation, insertion, etc. Refer to Table 3-4 for detailed descriptions of the types of texts selected with their associated characteristics in these three perspectives.

Language-specific issues in measuring listening ability

Issues regarding the standard dialect (i.e., standard language) in different speech communities represent a particular challenge for DLPT5 test developers. The ILR offers no guidance about languages in which many standard dialects are associated with one single language. Similarly there are no ability statements in the ILR for such situations. French and Spanish offer examples of speech communities located in widespread geographic and geopolitical regions that may have a different perception of what constitutes “standard language.” As concerns these languages, stakeholders tasked DLPT5 test developers to sample from a wide-range of standard variants. In another case, stakeholders have directed DLPT5 test developers to develop separate tests in Lusitanian and Brazilian Portuguese.

The lack of guidance within the ILR Skill Level Descriptions as regards languages like French and Spanish is problematic for construct definition, test development, and score interpretation. It is difficult to determine what listening ability is in light of the multiple Spanish or French “standards,” what listeners at various levels (especially at the lowest levels) are able to process, and what listening texts are to be included for appropriate representation of the target language use context(s).

Modern Standard Arabic (MSA) and its regional variants pose another challenge. MSA is the means of communication in writing in formal (i.e., social, educational, and/or religious) contexts. It may also play a role in speaking, especially when Arabic speakers from different nations or dialect regions, e.g., Egypt, Morocco, the Levant, etc., want to speak to one another. MSA is the macrolanguage that helps bridge the gap between the dialects of Arabic. However, this spoken MSA develops variants which result from the pronunciation and suprasegmental influences of the speakers' home regions and vernaculars. As in the case of the Spanish and French DLPT5s, an executive decision was made to sample the different variants of spoken MSA in the Arabic world.

The interplay between MSA and the Arabic dialects may also have an effect on measurement. It is unclear to what extent, if any, knowledge of MSA may influence scores for examinees who take a test in one of the Arabic dialects. Regarding this issue, it was decided that the developers of Arabic-dialect DLPT5s should adhere to the guidelines concerning skills measured at each ILR level and should include a representative sampling of spoken dialect texts. Although these texts likely contained, to varying degrees, loan words from MSA, it was decided that test takers who were able to demonstrate the ability specified in the ILR descriptor should be awarded commensurate credit when taking an Arabic-dialect test.

One issue remains in terms of testing dialects or vernaculars that are not widely spoken, and that is the matter of availability of fully authentic materials. In the development of the tests for Cebuano, Chavacano, and Tausug, the DLPT5 TL test developers were constrained to purpose-write a larger-than-normal percentage of the listening texts (especially at the lowest ILR levels) due to a sheer lack of resources. In doing so, they did, however, make every attempt to maintain the characteristics appropriate for the target-language use context and to ensure that native target-language users would regard these purpose-written texts as acceptable.

Question types

The types of questions and the information requested affect listening comprehension. Freedle and Kostin (1999) studied the interaction between one TOEFL listening text type, mini-talk, and task difficulty. Their finding suggested that the topic and the discourse structure affect difficulty but the degree of lexical overlap and the location of the necessary information to answer the questions are the two most important determinants in item difficulty. When the necessary information to answer the question comes near the beginning of the text, or when it is repeated, the items become easier. Lexical overlap refers to words that are used in the passages and also found in the questions or the options. The finding of lexical overlap may not be applicable to the DLPT context since TOEFL is a monolingual test. However, translation of lexical items may play a similar role and needs to be investigated.

Shohamy and Inbar (1991) consider how type of question influences success in second-language listening comprehension tests. They found that subjects perform better on questions answerable by referring to local cues in the text than on those answerable by referring to global cues. In other words, it is more difficult to generalize, infer and synthesize information than to look for specific information. These findings are reflected in the task types used at the different ILR levels (see table 3-4).

Questions eliciting explicit information are easier than questions asking for implicit information, which may also be conveyed through variables found in the spoken language such as stress and intonation (Bejar et al., 2000). Questions are easier when they require concrete information than abstract information. Questions that require recall of exact information are easier than questions that require extracting the main ideas. Finally, questions that require processing of less information are easier than questions that require processing more information (Buck, 2001). These findings are reflected in task-development practices used in the DLPT5.

The DLPT5 listening questions have two broad types: comprehension of literal meaning and comprehension of inferential meaning. Questions of literal comprehension require explicit information. At lower levels, examinees are required to process concrete information in order to answer the questions. The information required may be the exact information as delivered in the passage for examinees to identify or scattered throughout the texts, which requires examinees to integrate and generalize. At higher levels, examinees are expected to process abstract information and answer questions that require generalization and synthesis of information. Questions of inferential comprehension require examinees to process implicit and abstract information, which requires generalization, synthesis and inference of the listening input. These requirements can pose a special challenge in listening comprehension texts: the listener must synthesize elements of meaning that are no longer present, since the passage is processed in real time and no transcript is provided; and the listener must extract meaning from cues such as tone of voice or various other vocal modulations. We think these two question types cover the purposes of what the DLPT5 Listening Test measures; that is, listening for specific information, listening for basic comprehension, and listening to learn. These two question types also tap the range of sub-skills associated with the levels of listening proficiency. Refer to Table 3-4 for the list of sub-skills measured.

One very important inference that can be made from almost all of the findings discussed above concerns difficulty. Although there is a general trend for texts and tasks to become more difficult as ILR level increases, empirical difficulty (as measured by various statistical models) and ILR level are not perfectly correlated. Moreover, because of the complex interplay of factors discussed above, there is often substantial variation in difficulty within the texts and tasks at a particular ILR level for individual examinees. Hence, we expect to see a range of difficulty for tasks at a given ILR level.

- Describing listening abilities, texts and question types

In this section, we will list all the required listening sub-skills, describe the listening texts and tasks, and explain the processing demand required of the examinees for successful listening.

Table 3-4 presents detailed descriptions of the targeted listening sub-skills, types of texts, tasks and questions from ILR Level 1 – Level 4. Six sub-headings are provided:
ILR Skill Level Descriptions: the targeted ILR skill level descriptions are provided verbatim,
Description of Expected Ability: the abilities most salient at the given ILR level are extrapolated and listed,
Skills to Be Assessed: those salient abilities realized in terms of measureable skills are listed;

Target Language Input: detailed descriptions regarding the characteristics of the spoken texts for each of the ILR levels (e.g., text type, length, linguistic complexity, discourse organization, sound quality, etc.) are provided,
Tasks/Questions and Expected Responses: a description of the item types and the focuses of the questions is listed, and
Cognitive Load: a description of the expected processing demand on the part of the examinees in order to be successful at the given ILR level is provided.

The content of the DLPT5 Listening Test includes areas and topics covered in the Final Learning Objective (FLO). Refer to Table 2-1 for a description of examples of possible content areas within the FLO. The channel of input (i.e., listening passages in the target language) is audio.

Table 3-4

ILR Level	**Listening 10**
ILR Skill Level Descriptions	Sufficient comprehension to understand utterances about basic survival needs and minimum courtesy and travel requirements in areas of immediate need or on very familiar topics, can understand simple questions and answers, simple statements and very simple face-to-face conversations in a standard dialect. These must often be delivered more clearly than normal at a rate slower than normal with frequent repetitions or paraphrase (that is, by a native used to dealing with foreigners). Once learned, these sentences can be varied for similar level vocabulary and grammar and still be understood. In the majority of utterances, misunderstandings arise due to overlooked or misunderstood syntax and other grammatical clues. Comprehension vocabulary inadequate to understand anything but the most elementary needs. Strong interference from the candidate's native language occurs. Little precision in the information understood owing to the tentative state of passive grammar and lack of vocabulary. Comprehension areas include basic needs such as: meals, lodging, transportation, time and simple directions (including both route instructions and orders from customs officials, policemen, etc.). Understands main ideas. (Has been coded L-1 in some nonautomated applications.) [Data Code 10]
Description of Expected Ability	Able to understand: • Main ideas • Explicitly stated essential information
Skills to Be Assessed	Non-participatory listening in the ability to understand • main ideas and • explicitly stated essential information
Target Language Input	
Mode / Purpose	Orientation (Child) – the purpose of listening at this level is to fulfill basic needs such as meals, lodging and transportation, time and simple directions, minimum courtesy and travel requirements. Listeners listen for basic and factual information.

Speech Types / Types of Spoken Texts	• Simple, short transactional exchanges e.g., service encounters at stores, hotels, etc. • Simple and/or routine conversations, e.g., between friends, family members, colleagues. • Simple announcements or messages • Simple and short instructions or directions
Degree of planning	Speech segments at this level may be planned as in an announcement of a public event, or spontaneous as in business or interpersonal exchanges
Length	Up to 45 seconds
Syntactic Complexity	Utterances consist of structures representing the most basic or formulaic sentence or phrase types of the target language
Lexical Range	Words of concrete nature relevant to immediate and/or basic needs
Discourse Structure	Includes simple, short formal and informal speech segments. Utterances are literal and information is conveyed in a straightforward manner. Utterances may contain repetitions, redundancies and/or paraphrase. Discourse structure at this level is highly predictable and formulaic.
Sound quality	Clear enough that listeners are able to hear all utterances in the given speech segment with or without background noise.
Speech rate	Speed of utterances ranges from slower than normal to normal rate of speech as judged by native speakers of the target language.
Dialects and Accents	Utterances are delivered in a standard dialect and accent as judged by native speakers of the target language.
Other paralinguistic features	Utterances display oral characteristics found in the target language, such as fillers, pauses, hesitations, repairs, omissions, etc.
Sources	Speech segments may be obtained from the real world such as the Internet, TV/ radio broadcasts, recordings of real encounters, etc. Speech segments may also be created in-house or re-recorded in the studio if authentic speech segments do not reflect the types of linguistic features or topical areas considered to be at level.
Focus of Task/Question	- Main idea - Explicitly stated essential information
Cognitive Load	Tasks at this level require examinees to process explicit and concrete information. Test takers read the questions listen to the passage once. They are required to remember the necessary information and either identify the correct option in MC or provide the correct answer in CRT.

ILR Level	Listening 16
ILR Skill Level Descriptions	Sufficient comprehension to understand short conversations about all survival needs and limited social demands. Developing flexibility evident in understanding a range of circumstances beyond immediate survival needs. Shows spontaneity in understanding by speed, although consistency of understanding is uneven. Limited vocabulary range necessitates repetition for understanding. Understands more common time forms and most question forms, some word order patterns, but miscommunication still occurs with more complex patterns. Cannot sustain understanding of coherent structures in longer utterances or in unfamiliar situations. Understanding of descriptions and the giving of precise information is limited. Aware of basic cohesive features (e.g., pronouns, verb inflections) but many are unreliably understood, especially if less immediate in reference. Understanding is largely limited to a series of short, discrete utterances. Still has to ask for utterances to be repeated. Some ability to understand facts. (Has been coded L-1+ in some nonautomated applications.) [Data Code 16]
Description of Expected Ability	Able to understand: • Main ideas • Explicitly stated important/essential information
Skills to Be Assessed	Non-participatory listening in the ability to understand • main ideas and • explicitly stated essential information
Target Language Input	
Mode / Purpose	Mixed Orientation / Instructive Mode (Child) – the purpose of listening at this level is to fulfill basic needs and limited social and workplace demands. Listeners listen for factual information.
Speech Types / Types of Spoken Texts	• Simple, short transactional exchanges e.g., service encounters at stores, hotels, etc. • Simple and/or routine conversations, e.g., between friends, family members, colleagues. • Simple announcements or messages • Simple and short instructions or directions • Simple and short narration/description of events, places, or people, etc.
Degree of planning	Speech segments at this level may be planned, as in an announcement of a public event, or spontaneous, as in transactional or interpersonal exchanges.
Length	Up to 60 seconds
Syntactic Complexity	Utterances consist of basic time frames and structures representing

	simple or formulaic sentence or phrase types of the target language.
Lexical Range	Words of a concrete nature relevant to basic needs
Discourse structure	Utterances are literal but may contain pragmatic aspects of language use, e.g., "Could you open the door?" with the intended meaning of "Open the door, please". Utterances may contain repetitions, redundancies and/or paraphrase. Discourse structure at this level is predictable.
Sound quality	Clear enough that listeners are able to hear all utterances in the given speech segment with or without background noise.
Speech rate	Speed of utterances ranges from slower than normal to normal rate of speech as judged by native speakers of the target language.
Dialects and Accents	Utterances are delivered in a standard dialect and accent as judged by native speakers of the target language.
Other paralinguistic features	Utterances display oral characteristics found in the target language, such as fillers, pauses, hesitations, repairs, omissions, etc.
Sources	Speech segments may be obtained from the Internet, TV/ radio broadcasts, recordings of real encounters, etc. Speech segments may also be semi-scripted and recorded in the studio.
Focus of Task/Question	- Main idea - Explicitly stated essential information
Cognitive Load	Tasks at this level require examinees to process explicit and concrete information. Test takers read the questions listen to the passage once. They are required to remember the necessary information and either identify the correct option in MC or provide the correct answer in CRT.

ILR Level	Listening 20
ILR Skill Level Descriptions	Sufficient comprehension to understand conversations on routine social demands and limited job requirements. Able to understand face-to-face speech in a standard dialect, delivered at a normal rate with some repetition and rewording, by a native speaker not used to dealing with foreigners, about everyday topics, common personal and family news, well-known current events and routine office matters through descriptions and narration about current, past and future events; can follow essential points of discussion or speech at an elementary level on topics in his/her special professional field. Only understands occasional words and phrases of statements made in unfavorable conditions, for example through loudspeakers outdoors. Understands factual content. Native language causes less interference in listening comprehension. Able to understand facts; i.e., the lines but not between or beyond the lines. (Has been coded L-2 in some nonautomated applications.) [Data Code 20]
Description of Expected Ability	Able to understand facts / factual content in terms of: • Main ideas • Major details or important supporting information • Sequence of events • Causal and effect
Skills to Be Assessed	Non-participatory listening in the ability to understand • main ideas, • explicitly stated major details, or important supporting information • sequence of events • cause and effect
Target Language Input	
Mode / Purpose	Instructive – the purpose of listening at this level is to obtain factual information to fulfill routine social or workplace demands. Examinees may be required to listen for specific information.
Speech Types / Types of Spoken Texts	Contain formal or informal speech samples from the following: • conversations on routine social demands, • exchanges related to workplace requirements of a limited nature, • descriptions / narrations about events of a concrete nature, or • instructions or directions
Degree of planning	Speech segments at this level may be planned, as in a news report, or spontaneous, as in daily conversations.

Length	Up to 80 seconds
Syntactic Complexity	Utterances represent different time frames, i.e., simple past, present, and/or future tenses. Utterance structures may consist of simple or complex sentence structures or meaningful units in the target language.
Lexical Range	Utterances consist of words of a concrete nature, and may contain the most common idiomatic expressions
Discourse structure	Speech segments are literal and deal with concrete, factual information. They may contain repetitions, redundancies and/or paraphrase.
Sound quality	Utterances in the given speech segment may or may not have background noise. In segments with background noise, examinees are able to hear all utterances clearly.
Speech rate	Utterances are delivered at normal rate of speech as judged by native speakers of the target language.
Dialects and Accents	Utterances are delivered in a standard dialect and accent as judged by native speakers of the target language.
Other paralinguistic features	Utterances may display oral characteristics found in the target language, such as fillers, pauses, hesitations, repairs, omissions, etc.
Sources	The majority of speech segments at this level are obtained from authentic, real-time sources such as the Internet, TV/ radio broadcasts/programs, recordings of real encounters, etc. Authentic speech samples may undergo minimum editing to improve the sound quality and make the length fit the requirement. Some speech segments may be created in-house or recorded in the studio.
Focus of Task/Question	- Main idea - Major details or important supporting information - Sequence of events - Cause and effect
Cognitive Load	Tasks at this level require examinees to use their knowledge of how the target language culture functions at a concrete level and process explicitly stated factual information presented in the rhetorical structures characteristic of the listening text types at this level. Examinees are required to remember ad integrate factual information and identify the correct option in MC items or provide the correct answers in CRT items.

ILR Level	Listening 26
ILR Skill Level Descriptions	Sufficient comprehension to understand most routine social demands and most conversations on work requirements as well as some discussions on concrete topics related to particular interests and special fields of competence. Often shows remarkable ability and ease of understanding, but under tension or pressure may break down. Candidate may display weakness or deficiency due to inadequate vocabulary base or less than secure knowledge of grammar and syntax. Normally understands general vocabulary with some hesitant understanding of everyday vocabulary still evident. Can sometimes detect emotional overtones. Some ability to understand implications. (Has been Coded L-2+ in some nonautomated applications.) [Data Code 26]
Description of Expected Ability	Able to understand facts / factual content in terms of: • Main ideas • Supporting details • Sequence of events Can sometimes detect emotional overtones and understand implications
Skills to Be Assessed	Non-participatory listening in the ability to • understand main ideas, • understand explicitly stated supporting details, • understand sequence of events, • understand cause and effect • draw simple inferences.
Target Language Input	
Mode / Purpose	Mixed Instructive / Evaluative Mode (Child)– the purpose of listening at this level is to obtain information in order to deal with routine social and workplace demands as well as to follow simple discussions on general/concrete topics.
Speech Types / Types of Spoken Texts	Contain formal or informal speech samples from the following: • conversations and/or workplace exchanges • descriptions / narrations about events, places or people, • explanations, instructions or directions, and/or • discussions/lengthy exchanges on concrete topics such as weather, traffic, etc.
Degree of planning	Speech segments at this level may be planned, as in an announcement of a public event, somewhat planned, as in giving somewhat complex directions, or spontaneous, as in business or interpersonal exchanges.

Length	Up to 100 seconds
Syntactic complexity	Utterances represent different time frames, i.e., simple past, present, and/or future tenses. Utterance structures may consist of simple sentence structures or meaningful units in the target language.
Lexical Range	Utterances mostly consist of words of concrete nature and idiomatic expressions to fulfill daily demands. Words of abstract nature may appear in segments in which the speakers express emotions or personal view.
Discourse structure	Speech segments may go beyond literal meanings and take on (emotional) overtones. Utterances may contain repetitions, redundancies and/or paraphrase.
Sound quality	Listeners are able to hear most utterances in the given speech segment with or without background noise. In segments with background noise, the degree of background noise does not interfere with comprehension of those segments.
Speech rate	Utterances are delivered at normal rate of speech as judged by native speakers of the target language.
Dialects and Accents	Utterances are delivered in a standard dialect and accent considered by native speakers of the target language.
Other paralinguistic features	Utterances display oral characteristics found in the target language, such as fillers, pauses, hesitations, repairs, omissions, etc.
Sources	Speech segments may be obtained from the Internet, TV/ radio broadcasts, recordings of real encounters, etc. Authentic speech segments may undergo minimum editing. Speech segments may also be created and recorded in the studio.
Focus of Task/Question	- Main ideas - Supporting details - Sequence of events - Cause and effect - Simple inferences
Cognitive Load	Tasks at this level require examinees to use their knowledge of how the target language culture functions at a concrete level, and process explicitly stated factual information presented in the rhetorical structures characteristic of the listening text types at this level. Examinees are required to remember and integrate factual information and identify the correct option in MC items or provide the correct answers in CRT items. For some passages and questions, examinees are required to integrate, summarize the information and express judgments about the information they heard.

ILR Level	Listening 30
ILR Skill Level Descriptions	Able to understand the essentials of all speech in a standard dialect including technical discussions within a special field. Has effective understanding of face-to-face speech, delivered with normal clarity and speed in a standard dialect on general topics and areas of special interest; understands hypothesizing and supported opinions. Has broad enough vocabulary that rarely has to ask for paraphrasing or explanation. Can follow accurately the essentials of conversations between educated native speakers, reasonably clear telephone calls, radio broadcasts, news stories similar to wire service reports, oral reports, some oral technical reports and public addresses on non-technical subjects; can understand without difficulty all forms of standard speech concerning a special professional field. Does not understand native speakers it they speak very quickly or use some slang or dialect. Can often detect emotional overtones. Can understand implications. (Has been coded L-3 in some nonautomated applications.) [Data Code 30]
Description of Expected Ability	Able to • understand the essentials of all speech in a standard dialect • understand implications • detect emotional overtones
Skills to Be Assessed	Non-participatory listening in the ability to • understand major ideas that may be explicitly stated or be implicit in the text. • draw appropriate conclusions from a speaker's/speakers' remarks. • understand major points supporting a line of argumentation presented by the speaker(s). • understand different points of view presented by the speaker(s). • understand speaker-intended implications or inferences. • understand the significance of important cultural references or allusions cited by the speaker(s). • Identify the intent of the speaker(s).
Target Language Input	
Mode / Purpose	Evaluative Mode (Child) – listening at this level is to understand the essential points in the given speech samples either explicitly stated or implicitly expressed. Listeners are required to listen and integrate information presented in the listening text / speech sample.
Speech Types / Types of Spoken Texts	Formal or informal speech samples from the following: • Extended conversations and/or interviews on abstract or concrete

	but complex topics • Instructions / explanations of a technical nature • Commentaries • Discussions / debates • Lectures / public speeches
Degree of planning	Speech segments at this level may be planned, as in a news report, somewhat planned, as in a lecture, or spontaneous, as in discussions in a town hall meeting.
Length	Up to 120 seconds
Syntactic complexity	Utterances represent different time frames and aspects, i.e., simple past, present, and/or future tenses, present/past perfect and/or progressive. Utterance structures may consist of complex sentence structures or chunks of meaningful units in the target language.
Lexical Range	Utterances consist of a wide range of concrete or abstract words, as well as idiomatic expressions.
Discourse structure	Speech segments may demonstrate a variety of discourse structures and styles. The discourse organization of the speech segments may be structured as in an interview or casual as in a conversation. They may contain repetitions, redundancies and/or paraphrase.
Sound quality	Utterances in the given speech segment may come with background noise, interruptions, but comprehensible to the average native speakers of the target language.
Speech rate	Utterances are delivered at normal rate of speech as considered by the target language users, though some speech samples may be slightly faster than others.
Dialects and Accents	Utterances are delivered in a standard dialect and accent but may show considerable phonological modification and individual variations.
Other paralinguistic features	Utterances display oral characteristics found in the target language, such as fillers, pauses, hesitations, repairs, omissions, etc.
Sources	The speech segments at this level may be obtained from authentic sources such as the Internet, TV/ radio broadcasts/programs, recordings of real encounters, etc. Speech segments can be created in-house. Authentic speech segments may undergo minimum editing.
Focus of Task/Question	- Main idea - Supporting information - Sequence of events / cause and effect - Lines of argumentation in support of the speaker's view / speakers' views

	- Speaker intended implications / inferences / conclusions - Speaker's tone / attitude / position - Significance of a given idiomatic/colloquial expression
Cognitive Load	Tasks at this level require examinees to use their well developed language knowledge, and process abstract information on socio-political issues in areas where the target language is spoken. Examinees are required to remember, understand, analyze and integrate the information presented in the texts, and identify the correct option in MC items or provide the correct answers in CRT items.

ILR Level	Listening 36
ILR Skill Level Descriptions	Comprehends most of the content and intent of a variety of forms and styles of speech pertinent to professional needs, as well as general topics and social conversation. Ability to comprehend many sociolinguistic and cultural references. However, may miss some subtleties and nuances. Increased ability to comprehend unusually complex structures in lengthy utterances and to comprehend many distinctions in language tailored for different audiences. Increased ability to understand native speakers talking quickly, using nonstandard dialect or slang; however, comprehension is not complete. Can discern some relationships among sophisticated listening materials in the context of broad experience. Can follow some unpredictable turns of thought readily, for example, in informal and formal speeches covering editorial, conjectural and literary material in subject matter areas directed to the general listener. (Has been coded L-3+ in some nonautomated applications.) [Data Code 36]
Description of Expected Ability	Able to understand: • the content and intent of speech samples on general topics and/or related to workplace demands in a variety of forms and styles • the meaning and significance of the more common sociolinguistic and/or cultural references in the speech samples • speaker-intended implications, subtleties and/or nuances • speech samples tailored for difference audiences • fast speech • speech samples in a non-standard dialect and/or slang Able to follow unpredictable turns of thought
Skills to Be Assessed	Non-participatory listening in the ability to • understand major ideas that may be explicitly stated or be implicit in the text • draw appropriate conclusions from a speaker's remarks • understand major points supporting a line of argumentation presented by the speaker(s) • understand different points of views presented by the speaker(s) • understand implications conveyed by supra-segmental features (e.g., intonation, stress, etc.) • understand speaker-intended implications or inferences • understand the significance of the sociolinguistic and/or cultural references or allusions or slang used by the speaker(s) • understand fast speech samples • understand speech samples delivered in a non-standard dialect • identify the intent of the speaker(s)
Target Language	

Input	
Mode / Purpose	Mixed Evaluative / Projective Mode (Child) – listening at this level is to understand essential points given in complex speech samples either explicitly stated or implicitly expressed.
Speech Types / Types of Spoken Texts	Formal or informal speech samples from the following: • Extended conversations and/or interviews on abstract and/or concrete but complex topics related to FLO, • Commentaries • Discussions / debates • Lectures / public speeches
Degree of planning	Speech segments at this level may be planned, as in a presidential address, somewhat planned, as in a commentary, or spontaneous, as in discussions in a seminar/workshop
Length	Up to 140 seconds
Syntactic complexity	Utterances represent different time frames and aspects, i.e., simple past, present, and/or future tenses, present/past perfect and/or progressive. Utterance structures may consist of lengthy, unusually complex sentence structures or chunks of meaningful units in the target language.
Lexical Range	Utterances consist of a wide range of vocabulary, and idiomatic expressions with socio-cultural connotations. Utterances may also contain colloquial or slang expressions and/or non-standard language use.
Discourse structure	The discourse organization of the speech segments may be structured as in a formal interview or casual as in a conversation. Utterances may contain repetitions, redundancies and/or paraphrase but are self-contained.
Sound quality	Listeners are able to hear utterances in the given speech segment without background noise. In segment with background noise, the background noise may sometimes mask part of the utterances.
Speech rate	Speed of utterances ranges from normal rate to faster-than-normal rate of speech as considered by native speakers of the target language.
Dialects and Accents	Utterances may be delivered in a non-standard dialect and/or a non-standard accent as judged by native speakers of the target language. However, such utterances are understood by an average native speaker of the target language.
Other paralinguistic features	Utterances display oral characteristics found in the target language, such as fillers, pauses, hesitations, repairs, omissions, etc.
Sources	Speech segments may be obtained from the Internet, TV/ radio broadcasts, recordings of real encounters, etc. Speech segments may also be created in-house, for example, a debate on a complex social issue or a personal reflection on certain current affairs. Some authentic

	speech segments may undergo editing for length.
Focus of Task/Question	- Main idea - Supporting details - Lines of argumentation in support of the speaker's view / speakers' views - Speaker intended implications / inferences / conclusions - Speaker's tone / attitude / position - Significance of a given idiomatic/colloquial expression - Significance of a given sociolinguistic and/or cultural reference / allusion
Cognitive Load	Tasks at this level require examinees to use their well-developed language knowledge, and process abstract information on socio-political issues in areas where the target language is spoken. Examinees are required to remember, understand, analyze, integrate the information presented in the texts, express their judgments and opinions, and identify the correct option in MC items or provide the correct answers in CRT items.

ILR Level	Listening 40
ILR Skill Level Descriptions	Able to understand all forms and styles of speech pertinent to professional needs. Able to understand fully all speech with extensive and precise vocabulary, subtleties and nuances in all standard dialects on any subject relevant to professional needs within the range of his/her experience, including social conversations; all intelligible broadcasts and telephone calls; and many kinds of technical discussions and discourse. Understands language specifically tailored (including persuasion, representation, counseling and negotiating) to different audiences. Able to understand the essentials of speech in some non-standard dialects. Has difficulty in understanding extreme dialect and slang, also in understanding speech in unfavorable conditions, for example through bad loudspeakers outdoors. Can discern relationships among sophisticated listening materials in the context of broad experience. Can follow unpredictable turns of thought readily, for example, in informal and formal speeches covering editorial, conjectural and literary material in any subject matter directed to the general listener. (Has been coded L-4 in some nonautomated applications.) [Data Code 40]
Description of Expected Ability	Able to • understand all forms and styles of speech pertinent to social and professional needs in the standard dialect • understand fully the subtleties and nuances in speeches with extensive and precise vocabulary in the standard dialects on subjects relevant to the listener's professional needs within the range of his/her experience • understand the major points of speech in some non-standard dialects • understand the speaker / speakers' intent • understand language specifically tailored to different audiences • discern relationships among sophisticated listening materials in the context of broad experience • follow unpredictable turns of thought
Skills to Be Assessed	Non-participatory listening in the ability to • understand lines of argumentation by synthesizing information and ideas presented by the speaker(s). • understand major points supporting a speaker's argumentation / point of view. • draw appropriate conclusions or summarize. • understand subtle speaker-intended implications. • understand subtle argumentation. • place the speaker's message in a wider contextual framework.

	• understand register. • understand the speaker's attitude or tone. • understand colloquial and/or less frequently used idiomatic expressions. • understand the significance of the sociolinguistic and/or cultural references or allusions used by the speaker(s) • understand implications represented by supra-segmental features. • follow unpredictable turns of thought.
Target Language Input	
Mode / Purpose	Projective Mode (Child) – listening as this level is effortless as the listener is able to understand all varieties of speech and the different dialects used to deliver a given speech sample within the listener's world experience.
Speech Types / Types of Spoken Texts	Contain formal or informal speech for the general listener in the following: • Extended conversations and/or interviews on complex topics • Commentaries / editorials • Discussions / debates • Lectures / public speeches • Persuasion, negotiation, and/or counseling • Figurative speech / personal reflections • Speech sampled from different sub-culture groups / idiosyncratic speech samples
Degree of planning	Speech segments at this level may be planned, as in a political speech, somewhat planned, as in a panel discussion/debate, or spontaneous, as in conversations.
Length	Up to 160 seconds
Syntactic complexity	Utterances represent all time frames, aspects and moods. Utterance structures may consist of complex and compound sentence structures or chunks of meaningful units in the target language.
Lexical Range	Vocabulary used at this level - is extensive and/or precise, - may contain some low-frequency words, colloquialisms, high frequency slangs, regional use of idiomatic expressions, or idiosyncratic use of the target language.
Discourse structure	Speech segments demonstrate a variety of discourse structures and styles. The discourse organization of the speech segments may be structured as in a formal speech delivered in a convention or less formally, as in a conversation. Speech samples may contain repetitions,

	redundancies and/or paraphrase. Speech samples at this level also represent individualized speech, delivery styles and appropriate for the occasion.
Sound quality	Utterances in the given speech segment may contain background noise, interruptions, but comprehensible to average native speakers of the target language.
Speech rate	Utterances are delivered at normal to slightly faster rate of speech as judged by native speakers of the target language.
Dialects and Accents	Utterances are delivered in a standard dialect and/or non-standard dialect with a standard or non-standard accent containing considerable phonological modifications and individual variations.
Other paralinguistic features	Utterances display characteristics found in the target language, such as fillers, pauses, hesitations, repairs, omissions, etc.
Sources	The speech segments at this level should be obtained from authentic sources such as the Internet, TV/ radio broadcasts/programs, recordings of real encounters, etc. The speech segments can be prepared, pre-scripted and/or spontaneous.
Focus of Task/Question	- Main ideas / significance of the speaker's or speakers' message(s) in a wider contextual framework - Major points of argumentation supporting the speaker's view / speakers' views - Speaker-intended implications / inferences / conclusions - Speaker's tone / attitude / position - Significance of a given idiomatic/colloquial expression and/or a line of argumentation - Significance of a given cultural reference / allusion
Cognitive Load	Tasks at this level assume the examinees fully understand the content and context of the texts presented. Examinees are expected to synthesize the information. With their cultural knowledge as a facilitating factor, they are required to identify the correct option in MC items or provide the correct answers in CRT items.

3.4 Measuring sustained performance

The previous sections have outlined the operationalization of the ILR Skill Level Descriptions in the domains of passage selection and task development. As has been noted, however, the ILR is a holistic scale in which proficiency is demonstrated by sustained performance: each description of a skill level involves statements about the breadth of content or text types that language users can handle, as well as statements such as "generally," "typically," and "sometimes," which indicate that ILR level rests on a body of performance over a range of areas, not on performance

on a single text or task. This section considers how the DLPT5 system operationalizes the accuracy and breadth statements in the ILR to measure sustained performance at a level.

The underlying assumption behind the DLPT5 is that in order to demonstrate proficiency at a given level, examinees must show adequate comprehension on a variety of texts and tasks at that level. The two aspects that need to be further defined are "a variety" and "adequate comprehension." Regarding variety, the pools of items developed for DLPT5 tests are designed to sample as broadly as possible the range of texts found at each level in the real world, spanning the range of text types, content areas, and difficulty appropriate to the level. As has been noted above, each ILR level encompasses quite a wide range of all three of these characteristics. Regarding an adequate degree of comprehension, the DLPT5 criterion for demonstrating proficiency at a level is that an examinee should be able to answer correctly 70% of the questions in the pool at that level. In mastery testing, thresholds of 70-80% correct are often used to establish that an examinee has an adequate grasp of the subject being tested. For the DLPT5, the 70% criterion was set in consultation with a limited number of ILR experts, based on the notion that sustained performance means that language users at a given level can handle considerably more than half of the material at their level, but are not expected to understand all of it. [2]

The assumption, then, is that proficiency at a given level is defined as the ability to answer correctly 70% of a pool of questions that are representative of that level. For multiple-choice DLPT5s, there is a large pool of questions at each level, and passages are selected and tasks created such that the pool at each level represents the range of content, text types, and difficulty found at that level in the real world. Questions for the operational test forms are selected from this large pool. The questions on the operational test forms, however, do not necessarily form a representative sample of material at each level. Questions for the operational test forms are selected to provide as great a range as possible of content and text types; however, for practical reasons regarding the number of variables that can be brought into play in item selection, the difficulty of the items on the operational test forms does not always reflect the difficulty range of the items in the pool.

Difficulty statistics are obtained for questions in the large pool through administering the questions to a large group of examinees (see section 4 for additional information). The method of setting cut scores takes difficulty into account, so that the proportion of items examinees must answer correctly on the operational tests in order to receive a particular proficiency level score does not necessarily reflect an expectation that they will answer 70% of the questions on the operational test form at that level. For example, if the test developers select questions for an operational test form that are easier, on average, than the average difficulty of questions in the larger pool, examinees will be expected to get more than 70% of those questions correct. See Appendix B for a detailed description of how the calibration is performed and how cut scores are set. Note that Appendix B was written in 2007, originally for a different purpose. Nevertheless, the information is accurate and applies to the calibration methods for current DLPT5 tests.

[2] The use of the 70% criterion was based on agreement with NSA (Dr. Pardee Lowe) and the DLIFLC test developers. It will be necessary to establish the justifications for this figure more formally. The 70% figure has been arrived at through experience with tests very different from the DLPT5, especially with regard to the use of authentic materials, the mode of delivery, and item development techniques. DLIFLC is proposing that CASL study the issues and develop research-based definitions of mastery criteria and thresholds.

For constructed-response DLPT5s, there is no large pool of questions; each operational test form is made up of its own, smaller, pool of questions at each ILR level. Passages and questions for each operational test form at each level are selected to reflect, as much as possible, the range of content, text types, and difficulty of material at that level seen in the real world. Examinees must answer correctly approximately 70% of the questions at a given level in order to be awarded a score at that level.

4. The DLPT5 development process, quality assurance, and DLPT5 calibration

This section outlines the steps involved in DLPT5 development and the processes involved in ensuring that the tests operationalize the ILR properly.

4.1 Personnel selection and training

When initiating a test development project, the Test Development Division (TD) issues an announcement to recruit target language (TL) test developers, frequently referred to with the abbreviation 'TLEs,' from the staff at DLIFLC. In cases where TL resources are scarce at DLIFLC, TD may recruit TLEs from external as well as internal sources. TD will select, at a minimum, two TLEs for a project although, if resources are available, a DLPT5 project may sometimes consist of three TLEs. In recruiting TLEs, TD seeks applicants who have TL skills that equal or closely approximate those of a well-educated native user, and applicants are expected to have an ILR rating of Level 4 at a minimum. Applicants are also expected to be current in their use of the TL and exposure to the target-language use environment. TD attempts to find TLEs who have work experience that is expected to benefit the test development process, e.g., foreign language teaching or curriculum development experience, and who possess relevant educational qualifications, e.g., degrees in FL teaching, language testing, linguistics, or second language acquisition. A group of experienced test project managers, including the person who is slated to manage the project for which recruitment is taking place, reviews the applications submitted and, in most instances, interviews the applicants. After these steps, the project managers collectively make a recommendation to the Dean, TD, as to those applicants who would bring the strongest, most relevant skill sets to the project to be begun.

Once the TLEs have been formally selected and reassigned to duty in TD, a period of intensive training begins. Although the newly recruited TLEs may have already had some significant training in the ILR, for example in the form of training as oral proficiency interview (OPI) testers, they spend one week working with specialists in Test Review and Education (TRE), a division within the Directorate of Evaluation and Standardization (ESD) which specializes in matters related to interpreting the ILR level descriptors, placing texts and tasks on the ILR scale, and applying text typology when selecting passages. Not infrequently, the TD project manager attends such training with the newly selected TLEs in order to ensure that the test development team members and TRE's specialists are on the same wavelength in terms of ILR level assignment and of task appropriateness and text characteristics at the various ILR levels.

Following TRE's training in the ILR levels and text typology, the TLEs work intensively for 2-3 weeks with their project manager, selecting TL passages, rendering them into English (EN), and assigning ILR levels to those TL passages. (Passages are rendered into EN in order to give

access to the TL passages to individuals inside or outside the test development team who need to review the passages and items but do not know the TL in question.) One additional, critical focus of this part of the TLEs' training is to begin teaching them how to identify level-appropriate information in the TL texts and how to write English questions (i.e., tasks) that will elicit that information. In the case of multiple choice (MC) tests, TLEs begin learning the skills necessary to write effective distractor options. For tests in constructed response (CRT) format, TLEs learn how to prepare the key answer information that will become part of the scoring protocol for the test.

4.2 Item Development and Review Procedures for DLPT5

Throughout the two-year development cycle, TLEs are expected to work closely as a team, reviewing one another's work at every stage from passage selection, to ILR level assignment, through initial item development. This ensures that team members are fully informed and in agreement about the content of the test. Once these steps are completed, the TLEs and the project manager, as a group, review each passage and the item(s) developed to accompany it. This step allows the project manager to assess the progress of the TLEs in their fundamental work tasks, to provide them with ongoing mentoring about various aspects of their work, and to ensure consistency in terms of the quality of the item pool that the team will produce over the course of the project. In the case of a test in MC format, a team produces, at a minimum, a total of 360 test items. For a test in CRT format, a team produces, at a minimum, a total of 240 test items. These minimum numbers are dictated by the specifications for the relevant test formats.

In order to assure that DLPT5s, whether in multiple-choice or constructed-response format, operationalize the ILR as accurately as possible, the content of the tests undergoes multiple iterations of review outside of the test development team. The first level of review looks at the passage and set of test items as whole, focusing primarily on the appropriateness of the tasks elicited by the test items and the accuracy of the key answer information required to accomplish those tasks. This level of review is conducted by TD project managers who are critiquing and providing feedback on one another's work. The project manager whose passages and items are being reviewed discusses the input received with the project TLEs and incorporates changes as needed. The second level of review intends to ensure that the correct ILR level rating has been assigned to the passage and that the tasks posed in the questions also fit the assigned level. This review is conducted by the staff in TRE. Project managers in TD typically respond formally to the comments made by the TRE staff, and in the event that the TRE assessment of the ILR level does not match that of the test development team, the team and TRE must reach consensus on the ILR level. If such consensus cannot be reached, the passage and item set will not be included in the test. In a third level of review, trained, target-language raters outside of DLIFLC assess the test questions and key answer information in light of the target-language passages themselves. Beyond these three levels of review, in the event that a project manager is not a native speaker of English, all test items undergo review by a native speaker of English to ensure clarity and correctness. All test items must also be reviewed in their final form by the English editor within TD. Before releasing test materials for calibration, the test forms must undergo a global assessment by TRE to ensure that the content of the test forms is balanced in terms of topical breadth / distribution and relative difficulty. A TD-internal team also reviews the test forms to ensure that they meet formatting standards. The purpose of these various reviews is to ensure that

the test content of every DLPT5 conforms to the ILR Skill Level Descriptions and is as error-free as is possible prior to the start of test calibration/piloting.

4.3 Calibration of DLPT5 in multiple-choice format

DLPTs in multiple-choice format undergo calibration in order to gather item-level response data. (Please note that within DLIFLC, the word "validation" has been traditionally used to indicate the statistical calibration of test data.) The data are gathered from a broad a range of examinees at all ability levels in order to achieve as representative a sample as possible of target-language users. The responses gathered from this broad range of proficiency levels yield insights into the performance of items, and guide us in selecting items for the final operational test forms. For some languages, it is difficult to obtain enough validation examinees from within the operational testing population, due primarily to difficulties in releasing examinees from duty for the 20 or so hours required for validation testing. In such cases, DLPT5 test developers find examinees from other populations with some ability in the target language, including foreign nationals. Since DLPT5s are intended to measure global language proficiency of people who use English as their primary language in the workplace, we look for validation examinees who are either native speakers of English or target language users with very strong English skills[3].

For the purposes of calibration, validation examinees are encouraged to take both the reading and listening test components, but may choose to be tested in only one skill modality. They must respond to all the test questions for each skill modality in which they are tested. Prior to testing, examinees are provided with a DLPT5 Validation Familiarization Guide so that they can acquaint themselves with the test purpose and test process.

To assure that the calibration sample covers a wide a range of proficiency, an algorithm is used to assign each sample examinee a provisional ILR level based on his/her performance on the validation test, and to assess if the examinee test scores can be used for item analysis. The absolute minimum requirement is to obtain at least 10 examinees per provisional ILR level for the next stage, item analysis. The validation continues until the required number of examinees is obtained. Please note that the assignment of the provisional ILR levels to the validation population is for the test development team to track if sufficient data have been collected, and has no influence on the ILR level assigned using the operational tests, which is based on the result of statistical analysis (Please see below and Appendix B).

After data collection is completed, items are analyzed using both classical item-analysis techniques and item response theory (IRT). The test development team receives statistical information from a first analysis of all of the test items developed which allows it to examine how well the items have performed.

[3] The DLPT5 assumes English is the primary language in the target validation population, which consists of native speakers of English and speakers of other languages with very strong English skills. In selecting validation examinees, test developers have attempted to ensure that validation examinees have the requisite English skills. Many examinees are students at DLIFLC or universities or employees of government agencies; these are cases in which advanced proficiency in English is a requirement. In some other cases, a short screening test of English proficiency has been administered. Again given the time constraints of validation, it has not always been possible to administer the screening test.

Two types of information are used to determine item performance: item discrimination values and item difficulty values. Item discrimination is measured within classical item analysis using the point-biserial correlation. The point-biserial correlation is the correlation between the right/wrong score on a given item and the total score a test taker receives. It can have values from -1.0 to +1.0. A large positive point-biserial value indicates that test takers with high scores are getting a given item right and test takers with low test scores are getting the item wrong, i.e., the item is performing as expected. A low point-biserial value implies that test takers who get the item correct tend to do poorly on the test overall and test takers who get the item wrong tend to do well on the test overall, both of which are anomalous. Validation items with low or negative point-biserial correlations are not used to set cut scores. Such items are not selected for the operational test forms as they are not able to differentiate high-ability examinees from low-ability examinees.

Item difficulty is indicated within classical analysis as p-values. The p-values are the proportion of validation test takers that get an item correct. The p-value statistic ranges from 0 to 1. A high p-value indicates a given item is easy, and a low p-value indicates an item is difficult based on the test population from which the values are obtained.

Information on item discrimination and difficulty is also provided using the three-parameter logistic (3PL) model in Item Response Theory. The 3PL model is based on the idea that the probability of a correct response to a test item is a function of person and item parameters. The person parameter θ represents the ability of the individual (but we do not estimate θ for each examinee). The performance of an item is described by three item characteristics: item discrimination (parameter a), item difficulty (parameter b), and the probability of guessing (parameter c). The combined information is represented by an S-shaped curve called the item characteristic curve (ICC). The item characteristic curves (ICC) for the acceptable items at each ILR level are combined to form a test characteristic curve (TCC) at each ILR level. The EPC-Theta chart shown in Appendix B plots all of the ILR-based TCCs. It shows the expected proportion correct on each of the ILR item pools from a wide range of proficiency levels (i.e., from -4 to +4 on the θ). Together these TCC curves are used to assess the validity of the items. Appendix B provides a detailed description of the process used to construct the curves and how to interpret the values attached to them.

The use of the 3PL model to analyze the DLPT5 items makes it possible to develop more equivalent test forms and to move eventually to adaptive testing. It also provides information about the psychometric properties of individual test items that is more generalizable and sophisticated than that generated in classical analysis, since classical analysis is heavily reliant on the specific validation population. Please refer to Appendix B for a detailed description of the rationale underlying the model and item analysis process. Please also refer to Appendix B for a detailed description of the statistical methodology used to determine the validity, the reliability, and other important factors related to operational DLPT5 in multiple-choice format.

Using the classical and IRT statistical information, the test development team discards items that appear to be functioning poorly based on item discrimination (using the point-biserial correlation, IRT a-parameter, and analyses of how groups at different ability levels performed on each answer choice), item difficulty (using p-values in classical measurement and the IRT b-value,

comparing individual item difficulty to average difficulty for items at that level as a whole), or other factors such as the construct underlying an item. After removing those poorly functioning items from the pool, the test development team submits a selected pool of items for a second round of statistical analysis to verify that the item pool shows a clear differentiation of performance between the item pools at each level.

The same EPC-Theta chart that is used to assess the validity of the item pools is also used to determine the cut scores on the underlying proficiency (θ) scale. The proficiency criterion is that a person with threshold proficiency at a given ILR level can answer 70% of the pool items at that level correctly. Based on that criterion, the psychometrician determines the point at which the probability of a correct response for the items at a given ILR level is 70%; this point becomes the cut θ for that level. Thus θ cut-scores are established for each ILR level. Upon approval of the item pool for a given skill modality, the test development team selects the items that it proposes for inclusion in the two operational forms of the test for that skill. These proposed operational forms are again analyzed statistically to ensure that they are parallel and internally consistent. These cut scores are, in turn, used to generate the number-correct scoring tables for the operational test forms. Appendix B provides a detailed description of the process used to construct the curves and how to interpret the values attached to them.

Once it is established that the test forms for a given skill are statistically acceptable, the test development team assembles the required components, i.e., orientations, reading test passages, sound files with the listening passages, multiple-choice items, and key lists, and passes them on to be readied for computer delivery. Before the operational tests are released for the purpose of awarding official scores to examinees, the computer versions of the tests undergo one final review by native speakers of the target language and of English to ensure that the tests are as free of errors as is possible.

4.4 Piloting of DLPT5 in constructed-response format

Having completed test development and review, constructed-response tests undergo piloting. The purpose of piloting is to ensure that the passages and items in the test are functioning as expected. The test passages and items are divided into parallel pilot test forms and are administered to a sample of examinees. Since constructed-response tests are developed for languages where examinee populations are expected to be small, the number of examinees who can be tested in piloting will not be as large as those for multiple-choice tests. Nevertheless, an attempt is always made to administer the constructed-response pilot tests to as large a number of examinees and as a broad a range of examinees as is practicable. The test development team analyzes the written responses of the examinees from piloting in order to assess how examinees react to the questions asked and whether the key answer information required of examinees is appropriate. Based on its analysis, the test development team fine-tunes the test questions and/or key answer information required of examinees, as needed. The constructed-response test undergoes successive iterations of piloting and refinement based thereon until it is ready for to be used in its operational forms. The test development team assembles the required components, i.e., the orientations, reading test passages, sound files with the listening passages, and test questions, and passes them on to be readied for computer delivery. Before the operational tests are released for the purpose of awarding official scores to examinees, the computer versions of the tests undergo one final

review by native-speakers of the target language and of English to ensure that the tests are as free of errors as is possible.

The test development team must also assemble a Scoring Protocol for each form of the test. The scoring protocol includes for each passage in the test the orientation, the target-language passage itself, an English-language rendering of the target language, the test questions, the key answer information for each question, and a crediting scheme for awarding credit for each test question. The Scoring Protocol is also reviewed for accuracy and completeness before the test is put into use operationally.

4.5 CRT scorer training and scorer maintenance
A pool of in-house scorers scores all CRT tests. New constructed-response test scorers are competitively selected and have to go through a 2-day training workshop. An exit test is given to those who perform satisfactorily in the workshop. Those who pass the test are temporarily certified and are paired with a senior scorer each time they score CRT tests until they are fully certified. During the training period, a trainer independently scores the same sets of tests as the new scorer does in order to check the new scorer's rating performance. In addition, 10% of all scored CRT tests are rescored every month to monitor scorer performance. Those who give scores different from the final scores will be given individualized sessions to go over their scoring and scoring-related issues.

5. Test maintenance

The integrity of the DLPT5 testing system relies on test users' confidence in the tests. To ensure DLPT5 test validity and usability, standardized validation procedures are being put in place for ongoing evaluation of all current DLPT5 tests. DLPT5 test maintenance focuses on examining the test construct and encompasses three components for DLPT5 tests in the multiple-choice format: content evaluation, ILR evaluation, and statistical evaluation. Each of the operational DLPT5 multiple-choice tests undergoes an evaluation of all three components at regular intervals. If a given DLPT5 fails to meet the required criteria, the result will be either (1) modification of the test (through recalibration of items, resetting of cutscores, or elimination/revision/replacement of items) or (2) replacement of the test forms.

The content evaluation consists of examination in terms of content coverage, appropriateness of passages with their associated tasks, and how well the tasks function. The ILR evaluation verifies whether the individual passages and tasks conform to the ILR level skill descriptions as well as how well the test as a whole is able to tap the abilities specified in the ILR Skill Level Descriptions. Statistical evaluation examines item performance, compatibility of test forms, test performance over time, and whether current cut scores still represent ILR levels adequately.

For the DLPT5 tests in constructed-response format, in addition to the components mentioned above, regular scorer training will be conducted to ensure normed rating behavior. Rater behavior will further be monitored through statistical procedures.

Test maintenance is an ongoing process. Staff members in Test Development and the Test Review and Education Divisions are responsible for this crucial task.

6. Future directions

The section addresses a number of matters which have been mentioned about the DLPT5 in terms of how it works, how it might be improved, and how DLIFLC may want to deal with these matters.

6.1 Lower-ability examinees

One area of discussion has addressed testing lower-ability, examinees. Questions have been raised as whether DLPT5 test the lower level accurately enough and whether taking a full-range DLPT5 may demoralize examinees who must work through many items that are well beyond their level of ability.

One possible way to eliminate this problem would be the computer-adaptive, multi-stage delivery format. Such a test would consist of non-overlapping panels (independent sub-tests); examinees would typically take only one panel of the test. Each panel would consist of a number of sub-panels at varying levels, and examinees would be routed to different sub-panels in order to determine their score, which would be reported in terms of ILR level. The routing of an examinee to a particular Stage 2 sub-panel would depend on the performance of that examinee on the Stage 1 sub-panel; similarly, the Stage 3 sub-panel selected will depend on the performance at Stage 2. DLPT5s in computer-adaptive, multi-stage delivery format are being developed in a number of high-density languages, but because such tests required a very large pool of validated test items, it will be some time before tests in this format become available.

There exists another alternative that may address the question of testing lower-ability examinees. Instead of using a single linear test comprised of passages and items of increasing difficulty spanning a large number of ILR levels, there is the possibility of breaking down the tests into modules covering a more limited range of ILR levels. The goal of such modules would be to increase measurement accuracy at these levels and/or to decrease the test administration time. The Directorate of Evaluation and Standardization is investigating this possible testing alternative.

6.2 Sustained performance

The DLPT5 currently uses the measure 70% for defining sustained performance, with the exception of the Modern Standard Arabic test for which the criterion has recently been adjusted to 55% for Listening, 65% for Reading, pending further study. However, the question has been asked whether using 70% to define sustained performance across a representative sample of test items is the best choice. It might be possible to employ a standard-setting methodology to test the appropriateness of the 70% figure. Additional definitional work needs to be done as well to clearly determine whether “threshold” is the appropriate term to describe the sort of sustained performance represented by a 70% criterion. This is especially important since the target language abilities described by the ILR skill descriptions, represent a range or band, rather than a single point.

6.3 English use in the DLPT5

DLPT5 are bilingual tests, and the use of English used in the test has been mentioned as an area that may warrant investigation. DLPT test items, whether in multiple-choice or constructed-response format, are designed to be as easy to read and comprehend as possible. Nevertheless, if multiple-choice test takers are not native speaker of English, processing the English in the questions and answer options may add extra strain and have an influence on their scores. The constructed-response tests are somewhat similar in that non-native English speakers taking the test are asked to read and then answer in writing English-language test questions. In both of these scenarios, it is unclear how much of an impact the level of English in the tests has. This question of the interplay of examinee and English in the tests remains at the moment unanswered.

6.4 Note-taking

The literature is inconclusive regarding whether note-taking is valuable when taking listening comprehension tests, in some respects because many different types of tests have been examined. Note-taking is permitted on DLPT5 constructed-response tests, since examinees can type whatever they wish in the text boxes as the passage is playing. Note-taking is not permitted on the lower-range DLPT5 multiple-choice tests, but is permitted on the upper-range multiple-choice listening tests, at the request of a major stakeholder. It would be useful to conduct research on whether note-taking makes a significant difference on the DLPT5 lower-range multiple-choice test, as many stakeholders would prefer to be able to take notes.

6.5 Memory and passage length

Concerns have been raised by some stakeholders that passages, particularly listening passages, on the DLPT5 tests are too long, and that long passages test memory capacity more than language ability. There is some evidence that automatic processing, a component of language ability, is relevant to memory capacity, but it is not clear whether this component ought to be accessible at the lower levels of ability. There is also some evidence that when the length of passages is due in part to redundancy, examinees at lower levels find them easier, not harder, than shorter, less redundant passages. The effects of passage length on the construct need more examination.

6.6 Interaction with audio

On the listening comprehension tests, examinees may not stop and start the audio at will. Concerns have been raised that higher-level examinees become frustrated at having to listen to a passage twice, and lower-level examinees become lost when they cannot repeat a passage, or portions of it, multiple times. The issue of how multiple audio playings interact with item difficulty and the test construct has not been thoroughly examined. Concern has been expressed that giving examinees unlimited capability to stop and start the audio might lead to the test's measuring transcription ability rather than general language proficiency, and that for those test-takers who must use the language in real time, as opposed to through recordings, a test that allowed such manipulation would not measure the ability they need to have. More research is needed to determine what the effects of unlimited stopping and starting would be.

6.7 Difficulty and ILR level

One of the thorniest problems with testing based on the ILR is that each ILR skill level represents a range of ability, with different profiles within that range, yet tests of receptive skills

typically involve scoring based on whether examinees answered questions correctly. Since the difficulty of items at the same level varies significantly (as predicted by the ILR), and since the hardest items at a given level can sometimes be more difficult than the easiest items at the next level up, making the connection between examinee right/wrong scores and the ILR is not transparent. It would be useful to find a principled way to tease apart difficulty and ILR level.

6.8 Performance differences among subgroups of examinees

The DLPT5 is taken by a diverse population: examinees have a variety of educational backgrounds, native languages, language training, and job-related language performance requirements. For example, some examinees studied at the Defense Language Institute; others learned the language in college, and still others are native speakers of the target language. Some examinees are interrogators, while others' jobs primarily consist of listening to recorded material. It would be useful to examine item performance among different subgroups of the testing population. We would expect that some items would be significantly easier for some subgroups at a given ILR level than for others; however, it would be desirable to ensure that, taken as a whole, the items in the calibration pool, and on the individual test forms, were not skewed significantly to favor one group over another.

References

Alderson, J.C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Anderson, A. & Lynch, T. (1988). *Listening*. New York: Oxford University Press.

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

Bejar, I., Douglas, D., Nissan, S. & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper*. TOEFL Monograph Series MS-19. Princeton, NJ: Educational Testing Service.

Berne, J.E. (2004). Listening comprehension strategies: a review of the literature. *Foreign Language Annals* 37(4), 521-533.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge University Press.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Bygate, M. (1987). *Speaking*. New York: Oxford University Press.

Carrell, P.L. (1991) Second-language reading: Reading ability or language proficiency? *Applied Linguistics* 12, 159-179.

Carrell, P.L., Dunke., P.A., & Mollaun, P. (2002). *The effects of notetaking, lecture length and topic on the listening component of TOEFL 2000*. TOEFL Monograph Series MS-23. Princeton, NJ: Educational Testing Service.

Carver, R.P. (1982). Optimal rate of reading prose. *Reading Research Quarterly* XVIII(1), 56-88.

Carver, R.P. (1984). Rauding theory predictions of amount comprehended under different purposes and speed reading conditions. *Reading Research Quarterly* XIX(2), 205-218.

Chapelle, A., Enright, M.K., & Jamieson, J.M. (2008). Test score interpretation and use. In Chapelle, A., Enright, M.K., & Jamieson, J.M. (Eds.) *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge

Cheng, H. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals*, 37(4),

Cheuvront, S.L. (2004) The effects of notetaking on listening and reading comprehension: review of the literature. Unpublished paper. Evaulation and Standardization, Defense Language Institute Foreign Language Center.

Chiang, C.S. & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly* 26, 345-374.

Child, J.R. (1987) Language proficiency levels and the typology of texts. In H. Byrnes and M. Canale (eds.), *Defining and developing proficiency: Guidelines, implementations and concepts*. Lincolnwood, IL: National Textbook Co.

Child, J.R. (1998). Language skill levels, textual modes, and the rating process. *Foreign Language Annals*, 31(3), 381-391.

Child, J.R. (1999) Analysis of texts and critique of judgment. Retrieved on July 24, 2008 from http://dspace.wrlc.org/bitstream/1961/3455/9/gurt_1999_08.pdf

Clapham, C.M. (1996). *The development of IELTS: a study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

Clark, J.L.D. & Clifford, R.T. (1988). The FSI/ILR/ACTFL proficiency scales and testing techniques. *Studies in Second Language Acquisition* 10, 129-147.

Conrad, L. (1985) Semantic versus syntactic cues in listening comprehension. *Studies in Second Language Acquisition* 7, 59-72.

Dunkel, P. (1991). Listening in the native and second/foreign language: Toward an integration of research and practice. *TESOL Quarterly*, 25, 431-457.

Dunkel, P., Henning, G., & Chaudron, C. (1994). The assessment of a listening comprehension construct: A tentative model for testing specification and development. *Modern Language Journal*, 77(2), 180-191.

Edwards, A.L. (1996). Reading proficiency assessment and the ILR/ACTFL text typology: a reevaluation. *The Modern Language Journal* 80, 350-361.

Enright, M.K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedle, M. (2000). *TOEFL 2000 reading framework: A working paper*. TOEFL Monograph Series MS-17. Princeton, NJ: Educational Testing Service.

Flowerdew, J. (1994). Research of relevance to second language lecture comprehension—An overview. In J. Flowerdew (Ed.), *Academic listening* (7-29). New York: Cambridge University Press.

Freedle, R., & Kostin, I., (1999). Does the test matter in a multiple-choice test of comprehension? The case for the construct validity of TEOFL's minitalks. *Language Testing* 16(1), 2-32.

Goh, C.M. (2002) Exploring listening comprehension tactics and their interaction patterns. *System* 30, 185-206.

Graesser, A.C. & Brintton, B.K. (1996). Models of understanding text. New Jersey: Lawrence Erlbaum Associates.

Graham, S. (2006). Listening comprehension: the learners' perspective. *System* 34(2), 165-182.

Griffiths, R. (1990). Speech rate and NNS comprehension: a preliminary study in time-benefit analysis. *Language Learning* 40(3), 311-336.

Hale, G.A. (1988). Student major field and text content: interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing* 5(1), 49-61.

Henning, G. (1991). *A study of the effects of variation of short-term memory load, reading responses length, and processing hierarchy on TOEFL listening comprehension item performance.* TOEFL Research Report No. 33. Princeton, NJ: Educational Testing Service.

In'nami, Y. (2006). The effects of test anxiety on listening test performance. System 34(3), 317-340.

Johnston, J., & Doughty, C. (2006). M.2 Technical Report: listening sub-skills. CDRL:A021 DID:DI-MISC-80508A Contract Number: MDA904-03-C-0543

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing* 19(2), 193-220.

Long, R.L. (1990). What you don't know can't help you. *Studies in Second Language Acquisition* 12, 65-80.

Lowe, P., Jr. (1986). Proficiency: Panacea, Framework, Process? A reply to Kramsch, Schulz, and, particularly, to Bachman and Savignon. *Modern Language Journal* 70(4), 391-397.

Lowe, P., Jr. (1999). Evidence for the greater ease of use of the ILR language skill level descriptions for speaking. Retrieved on July 24, 2008 from http://dspace.wrlc.org/bitstream/1961/3455/5/gurt_1999_04.pdf

Lowe, P., Jr. (2006). Divining what isn't there: Crosswalked implicature in the ILR listening skill level descriptions. Presentation given at the East Coast Organization of Language Testers, October 13, 2006, George Washington University.

Lund, R. (1991) A comparison of second language listening and reading comprehension. *Modern Language Journal*, 75(2), 196-204.

Mandler, J.M. (1978). A code in the node: the use of a story schema in retrieval. *Discourse Processes* 1, 114-135.

Markham, P.L. (1988). Gender differences and the perceived expertness of the speaker as factors in ESL listening recall. *TESOL Quarterly*, 22, 397-406.

O'Malley, J.M., Chamot, A.U., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics* 10, 418-437

Rost, M. (1990). *Listening in language learning*. New York: Longman.

Rubin, J. (1994). A review of second language listening comprehension research. *Modern Language Journal* 78(2), 199-221.

Schmidt-Rinehart, B.C. (1994). The effects of topic familiarity on second language listening comprehension. *Modern Language Journal* 78(2), 169-178.

Shohamy, E. (1984). Does the testing method make a difference? The case of a reading comprehension. *Language Testing* 1(2), 202-220.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: the effect of text and question. Language Testing, 8, 23-40.

Ur, P. (1984). Teaching listening comprehension. New York: Cambridge University Press.

Van Patten, B. (1988). How juries get hung: problems with the evidence for a focus on form in teaching. *Language Learning* 28, 243-260.

Appendix A: Interagency Language Roundtable Language Skill Level Descriptions

Preface

The following descriptions of proficiency levels 0, 1, 2, 3, 4, and 5 characterize spoken-language use. Each higher level implies control of the previous levels' functions and accuracy. The designation 0+, 1+, 2+, etc. will be assigned when proficiency substantially exceeds one skill level and does not fully meet the criteria for the next level. The "plus-level" descriptions, therefore, are subsidiary to the "base-level" descriptions.

A skill level is assigned to a person through an authorized language examination. Examiners assign a level on a variety of performance criteria exemplified in the descriptive statements. Therefore, the examples given here illustrate, but do not exhaustively describe, either the skills a person may possess or situations in which he/she may function effectively.

Statements describing accuracy refer to typical stages in the development of competence in the most commonly taught languages in formal training programs. In other languages, emerging competence parallels these characterizations, but often with different details.

Unless otherwise specified, the term "native speaker" refers to native speakers of a standard dialect.

"Well-educated," in the context of these proficiency descriptions, does not necessarily imply formal higher education. However, in cultures where formal higher education is common, the language-use abilities of persons who have had such education is [sic] considered the standard. That is, such a person meets contemporary expectations for the formal, careful style of the language, as well as a range of less formal varieties of the language.

These descriptions may be further specified by individual agencies to characterize those aspects of language-use performance which are of insufficient generality to be included here.

Interagency Language Roundtable Language Skill Level Descriptions: Listening

Listening 0 (No Proficiency)

No practical understanding of the spoken language. Understanding is limited to occasional isolated words with essentially no ability to comprehend communication. (Has been coded L-0 in some nonautomated applications. [Data Code 00]

Listening 0+ (Memorized Proficiency)

Sufficient comprehension to understand a number of memorized utterances in areas of immediate needs. Slight increase in utterance length understood but requires frequent long pauses between understood phrases and repeated requests on the listener's part for repetition. Understands with reasonable accuracy only when this involves short memorized utterances or formulae. Utterances understood are relatively short in length. Misunderstandings arise due to ignoring or inaccurately hearing sounds or word endings (both inflectional and non-inflectional), distorting the original meaning. Can understand only with difficulty even such people as teachers who are used to speaking with non-native speakers. Can understand best those statements where context strongly supports the utterance's meaning. Gets some main ideas. (Has been coded L-0+ in some nonautomated applications.) [Data Code 06]

Listening 1 (Elementary Proficiency)
Sufficient comprehension to understand utterances about basic survival needs and minimum courtesy and travel requirements in areas of immediate need or on very familiar topics, can understand simple questions and answers, simple statements and very simple face-to-face conversations in a standard dialect. These must often be delivered more clearly than normal at a rate slower than normal with frequent repetitions or paraphrase (that is, by a native used to dealing with foreigners). Once learned, these sentences can be varied for similar level vocabulary and grammar and still be understood. In the majority of utterances, misunderstandings arise due to overlooked or misunderstood syntax and other grammatical clues. Comprehension vocabulary inadequate to understand anything but the most elementary needs. Strong interference from the candidate's native language occurs. Little precision in the information understood owing to the tentative state of passive grammar and lack of vocabulary. Comprehension areas include basic needs such as: meals, lodging, transportation, time and simple directions (including both route instructions and orders from customs officials, policemen, etc.). Understands main ideas. (Has been coded L-1 in some nonautomated applications.) [Data Code 10]

Listening 1+ (Elementary Proficiency, Plus)
Sufficient comprehension to understand short conversations about all survival needs and limited social demands. Developing flexibility evident in understanding a range of circumstances beyond immediate survival needs. Shows spontaneity in understanding by speed, although consistency of understanding is uneven. Limited vocabulary range necessitates repetition for understanding. Understands more common time forms and most question forms, some word order patterns, but miscommunication still occurs with more complex patterns. Cannot sustain understanding of coherent structures in longer utterances or in unfamiliar situations. Understanding of descriptions and the giving of precise information is limited. Aware of basic cohesive features (e.g., pronouns, verb inflections) but many are unreliably understood, especially if less immediate in reference. Understanding is largely limited to a series of short, discrete utterances. Still has to ask for utterances to be repeated. Some ability to understand facts. (Has been coded L-1+ in some nonautomated applications.) [Data Code 16]

Listening 2 (Limited Working Proficiency)
Sufficient comprehension to understand conversations on routine social demands and limited job requirements. Able to understand face-to-face speech in a standard dialect, delivered at a normal rate with some repetition and rewording, by a native speaker not used to dealing with foreigners, about everyday topics, common personal and family news, well-known current events and routine office matters through descriptions and narration about current, past and future events; can follow essential points of discussion or speech at an elementary level on topics in his/her special professional field. Only understands occasional words and phrases of statements made in unfavorable conditions, for example through loudspeakers outdoors. Understands factual content. Native language causes less interference in listening comprehension. Able to understand facts; i.e., the lines but not between or beyond the lines. (Has been coded L-2 in some nonautomated applications.) [Data Code 20]

Listening 2+ (Limited Working Proficiency, Plus)

Sufficient comprehension to understand most routine social demands and most conversations on work requirements as well as some discussions on concrete topics related to particular interests and special fields of competence. Often shows remarkable ability and ease of understanding, but under tension or pressure may break down. Candidate may display weakness or deficiency due to inadequate vocabulary base or less than secure knowledge of grammar and syntax. Normally understands general vocabulary with some hesitant understanding of everyday vocabulary still evident. Can sometimes detect emotional overtones. Some ability to understand implications. (Has been Coded L-2+ in some nonautomated applications.) [Data Code 26]

Listening 3 (General Professional Proficiency)
Able to understand the essentials of all speech in a standard dialect including technical discussions within a special field. Has effective understanding of face-to-face speech, delivered with normal clarity and speed in a standard dialect on general topics and areas of special interest; understands hypothesizing and supported opinions. Has broad enough vocabulary that rarely has to ask for paraphrasing or explanation. Can follow accurately the essentials of conversations between educated native speakers, reasonably clear telephone calls, radio broadcasts, news stories similar to wire service reports, oral reports, some oral technical reports and public addresses on non-technical subjects; can understand without difficulty all forms of standard speech concerning a special professional field. Does not understand native speakers it they speak very quickly or use some slang or dialect. Can often detect emotional overtones. Can understand implications. (Has been coded L-3 in some nonautomated applications.) [Data Code 30]

Listening 3+ (General Professional Proficiency, Plus)
Comprehends most of the content and intent of a variety of forms and styles of speech pertinent to professional needs, as well as general topics and social conversation. Ability to comprehend many sociolinguistic and cultural references. However, may miss some subtleties and nuances. Increased ability to comprehend unusually complex structures in lengthy utterances and to comprehend many distinctions in language tailored for different audiences. Increased ability to understand native speakers talking quickly, using nonstandard dialect or slang; however, comprehension is not complete. Can discern some relationships among sophisticated listening materials in the context of broad experience. Can follow some unpredictable turns of thought readily, for example, in informal and formal speeches covering editorial, conjectural and literary material in subject matter areas directed to the general listener. (Has been coded L-3+ in some nonautomated applications.) [Data Code 36]

Listening 4 (Advanced Professional Proficiency)
Able to understand all forms and styles of speech pertinent to professional needs. Able to understand fully all speech with extensive and precise vocabulary, subtleties and nuances in all standard dialects on any subject relevant to professional needs within the range of his/her experience, including social conversations; all intelligible broadcasts and telephone calls; and many kinds of technical discussions and discourse. Understands language specifically tailored (including persuasion, representation, counseling and negotiating) to different audiences. Able to understand the essentials of speech in some non-standard dialects. Has difficulty in understanding extreme dialect and slang, also in understanding speech in unfavorable conditions, for example through bad loudspeakers outdoors. Can discern relationships among sophisticated listening materials in the context of broad experience. Can follow unpredictable turns of thought

readily, for example, in informal and formal speeches covering editorial, conjectural and literary material in any subject matter directed to the general listener. (Has been coded L-4 in some nonautomated applications.) [Data Code 40]

Listening 4+ (Advanced Professional Proficiency, Plus)
Increased ability to understand extremely difficult and abstract speech as well as ability to understand all forms and styles of speech pertinent to professional needs, including social conversations. Increased ability to comprehend native speakers using extreme nonstandard dialects and slang, as well as to understand speech in unfavorable conditions. Strong sensitivity to sociolinguistic and cultural references. Accuracy is close to that of the well-educated native listener but still not equivalent. (Has been coded L-4+ in some nonautomated applications.) [Data Code 46]

Listening 5 (Functionally Native Proficiency)
Comprehension equivalent to that of the well-educated native listener. Able to understand fully all forms and styles of speech intelligible to the well-educated native listener, including a number of regional and illiterate dialects, highly colloquial speech and conversations and discourse distorted by marked interference from other noise. Able to understand how natives think as they create discourse. Able to understand extremely difficult and abstract speech. (Has been coded L-5 in some nonautomated applications.) [Data Code 50]

Interagency Language Roundtable Language Skill Level Descriptions: Reading

R-0: Reading 0 (No Proficiency)
No practical ability to read the language. Consistently misunderstands or cannot comprehend at all. [Data Code 00]

R-0+: Reading 0+ (Memorized Proficiency)
Can recognize all the letters in the printed version of an alphabetic system and high-frequency elements of a syllabary or a character system. Able to read some or all of the following: numbers, isolated words and phrases, personal and place names, street signs, office and shop designations. The above often interpreted inaccurately. Unable to read connected prose. [Data Code 06]

R-1: Reading 1 (Elementary Proficiency)
Sufficient comprehension to read very simple connected written material in a form equivalent to usual printing or typescript. Can read either representations of familiar formulaic verbal exchanges or simple language containing only the highest frequency structural patterns and vocabulary, including shared international vocabulary items and cognates (when appropriate). Able to read and understand known language elements that have been recombined in new ways to achieve different meanings at a similar level of simplicity. Texts may include descriptions of persons, places or things: and explanations of geography and government such as those simplified for tourists. Some misunderstandings possible on simple texts. Can get some main ideas and locate prominent items of professional significance in more complex texts. Can identify general subject matter in some authentic texts. [Data Code 10]

R-1+: Reading 1+ (Elementary Proficiency, Plus)
Sufficient comprehension to understand simple discourse in printed form for informative social purposes. Can read material such as announcements of public events, simple prose containing biographical information or narration of events, and straightforward newspaper headlines. Can guess at unfamiliar vocabulary if highly contextualized, but with difficulty in unfamiliar contexts. Can get some main ideas and locate routine information of professional significance in more complex texts. Can follow essential points of written discussion at an elementary level on topics in his/her special professional field.
In commonly taught languages, the individual may not control the structure well. For example, basic grammatical relations are often misinterpreted, and temporal reference may rely primarily on lexical items as time indicators. Has some difficulty with the cohesive factors in discourse, such as matching pronouns with referents. May have to read materials several times for understanding. [Data Code 16]

R-2: Reading 2 (Limited Working Proficiency)
Sufficient comprehension to read simple, authentic written material in a form equivalent to usual printing or typescript on subjects within a familiar context. Able to read with some misunderstandings straightforward, familiar, factual material, but in general insufficiently experienced with the language to draw inferences directly from the linguistic aspects of the text. Can locate and understand the main ideas and details in material written for the general reader. However, persons who have professional knowledge of a subject may be able to summarize or perform sorting and locating tasks with written texts that are well beyond their general

proficiency level. The individual can read uncomplicated, but authentic prose on familiar subjects that are normally presented in a predictable sequence which aids the reader in understanding. Texts may include descriptions and narrations in contexts such as news items describing frequently occurring events, simple biographical information, social notices, formulaic business letters, and simple technical material written for the general reader. Generally the prose that can be read by the individual is predominantly in straightforward/high-frequency sentence patterns. The individual does not have a broad active vocabulary (that is, which he/she recognizes immediately on sight), but is able to use contextual and real-world cues to understand the text. Characteristically, however, the individual is quite slow in performing such a process. Is typically able to answer factual questions about authentic texts of the types described above. [Data Code 20]

R-2+: Reading 2+ (Limited Working Proficiency, Plus)
Sufficient comprehension to understand most factual material in non-technical prose as well as some discussions on concrete topics related to special professional interests. Is markedly more proficient at reading materials on a familiar topic. Is able to separate the main ideas and details from lesser ones and uses that distinction to advance understanding. The individual is able to use linguistic context and real-world knowledge to make sensible guesses about unfamiliar material. Has a broad active reading vocabulary. The individual is able to get the gist of main and subsidiary ideas in texts which could only be read thoroughly by persons with much higher proficiencies. Weaknesses include slowness, uncertainty, inability to discern nuance and/or intentionally disguised meaning. [Data Code 26]

R-3: Reading 3 (General Professional Proficiency)
Able to read within a normal range of speed and with almost complete comprehension a variety of authentic prose material on unfamiliar subjects. Reading ability is not dependent on subject matter knowledge, although it is not expected that the individual can comprehend thoroughly subject matter which is highly dependent on cultural knowledge or which is outside his/her general experience and not accompanied by explanation. Text-types include news stories similar to wire service reports or international news items in major periodicals, routine correspondence, general reports, and technical material in his/her professional field; all of these may include hypothesis, argumentation and supported opinions. Misreading rare. Almost always able to interpret material correctly, relate ideas and "read between the lines," (that is, understand the writers' implicit intents in text of the above types). Can get the gist of more sophisticated texts, but may be unable to detect or understand subtlety and nuance. Rarely has to pause over or reread general vocabulary. However, may experience some difficulty with unusually complex structure and low frequency idioms. [Data Code 30]

R-3+: Reading 3+ (General Professional Proficiency, Plus)
Can comprehend a variety of styles and forms pertinent to professional needs. Rarely misinterprets such texts or rarely experiences difficulty relating ideas or making inferences. Able to comprehend many sociolinguistic and cultural references. However, may miss some nuances and subtleties. Able to comprehend a considerable range of intentionally complex structures, low frequency idioms, and uncommon connotative intentions, however, accuracy is not complete. The individual is typically able to read with facility, understand, and appreciate contemporary

expository, technical or literary texts which do not rely heavily on slang and unusual items. [Data Code 36]

R-4: Reading 4 (Advanced Professional Proficiency)
Able to read fluently and accurately all styles and forms of the language pertinent to professional needs. The individual's experience with the written language is extensive enough that he/she is able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references. Able to "read beyond the lines" (that is, to understand the full ramifications of texts as they are situated in the wider cultural, political, or social environment). Able to read and understand the intent of writers' use of nuance and subtlety. The individual can discern relationships among sophisticated written materials in the context of broad experience. Can follow unpredictable turns of thought readily in, for example, editorial, conjectural, and literary texts in any subject matter area directed to the general reader. Can read essentially all materials in his/her special field, including official and professional documents and correspondence. Recognizes all professionally relevant vocabulary known to the educated non-professional native, although may have some difficulty with slang. Can read reasonably legible handwriting without difficulty. Accuracy is often nearly that of a well-educated native reader. [Data Code 40]

R-4+: Reading 4+ (Advanced Professional Proficiency, Plus)
Nearly native ability to read and understand extremely difficult or abstract prose, a very wide variety of vocabulary, idioms, colloquialisms and slang. Strong sensitivity to and understanding of sociolinguistic and cultural references. Little difficulty in reading less than fully legible handwriting. Broad ability to "read beyond the lines" (that is, to understand the full ramifications of texts as they are situated in the wider cultural, political, or social environment) is nearly that of a well-read or well-educated native reader. Accuracy is close to that of the well-educated native reader, but not equivalent. [Data Code 46]

R-5: Reading 5 (Functionally Native Proficiency)
Reading proficiency is functionally equivalent to that of the well-educated native reader. Can read extremely difficult and abstract prose; for example, general legal and technical as well as highly colloquial writings. Able to read literary texts, typically including contemporary avant-garde prose, poetry and theatrical writing. Can read classical/archaic forms of literature with the same degree of facility as the well-educated, but non-specialist native. Reads and understands a wide variety of vocabulary and idioms, colloquialisms, slang, and pertinent cultural references. With varying degrees of difficulty, can read all kinds of handwritten documents. Accuracy of comprehension is equivalent to that of a well-educated native reader. [Data Code 50]

Appendix B: Validity and Reliability of DLPT5 Multiple-Choice Tests

Validity and Reliability of DLPT5 Multiple-Choice Tests

J. Ward Keesling, Ph.D.

July 2007

Contents

List of Figures

List of Tables

Validity and Reliability of DLPT5 Multiple-Choice Tests

Introduction

This document discusses the procedures and guidelines that are used in the construction of operational forms of the DLPT5 multiple-choice tests. It also outlines and gives examples of the analyses used to establish the construct validity of the tests and the internal consistency reliability and parallel forms reliability of the operational tests. A separate document catalogues the reliability information for the tests as they are developed.

The discussion in this document begins with the gathering of item-level response data from a sample of examinees, and proceeds through the item analysis and calibration of the pool of items from which items are drawn for the operational forms. A part of this analysis addresses the construct validity of the items in the calibration pool. The discussion continues with the guidelines for selection of items for the operational forms, and the setting of cut scores. The discussion concludes with the computation of internal consistency reliability and parallel forms reliability indices for the operational forms.

For most multiple-choice test development projects in the DLPT5 series, two parallel forms are created for operational use in testing proficiency in each skill (listening and reading). Increasingly, however, there are languages for which listening tests in specific dialects are being developed, without a corresponding reading test (Iraqi Arabic, for example). In these cases two operational forms of listening proficiency DLPT5 are developed. Furthermore, there are projects to develop multiple-choice tests at levels of proficiency above level 3 on the Inter-Agency Language Roundtable (ILR) scale (specifically to assess proficiency at levels 3+ and 4) in select languages. These multiple-choice tests all use the methods described in this document.

Item-level response data

Item-level response data are gathered from a sample of examinees who are tested at DLIFLC, other military installations in the United States and overseas, and colleges and universities throughout the United States. We attempt to gather data on reading and listening from the same examinees, to economize on the expense of gathering the data. This means that an individual examinee will take between 400 and 600 items (covering both skills), broken up into sets of 50 or so items at a time. We try to allow for generous time limits so that there are no issues with speeding of the tests, so there may be as many as 10, 2-hour testing sessions for an individual examinee. Non-military examinees are paid for their participation; they are not given incentives for attaining high scores. Military examinees are encouraged to participate to ensure the soundness of the testing procedures that will later be used to determine whether or not they graduate from DLIFLC and receive proficiency pay.

We attempt to gather data from a broad sample of examinees, usually more than 100, with a deliberate attempt to obtain scores from subjects across all levels of proficiency. We do try to

limit participation of native and heritage speakers who do not have good command of the English language, but this is often difficult to assure because we do not routinely test English proficiency directly.[4]

The answer sheets from the tests are brought to DLIFLC for scanning, which includes procedures to check that the scanning was performed correctly and that the separate parts of the calibration test in a particular skill are all present for each examinee and assembled into one record for subsequent analysis.

Item analysis

The initial item analysis is performed with two statistical analysis packages: WINSTEPS (which calculates some classical item analysis statistics as well as Rasch model assessments of the quality of the items) and BILOG-MG (which computes additional classical item statistics and estimates the item parameters of the 3-parameter logistic analysis model). Other analysis packages are used by the project teams to examine the frequency of response choices.

The first phase of item analysis focuses on the identification of the pool of items that should be included in the calibration of items. Some items will have been flagged by BILOG-MG as unsuitable for calibration on the grounds that they have large negative biserial correlations with the total score. We impose the more rigorous restriction that no item with any negative degree of biserial or point-biserial correlation with total test score can be included in the calibration pool. In addition, test development teams are encouraged to eliminate items that have very low positive point-biserial correlations (below 0.10).

The analysis of response choices may also indicate reasons why some items should be eliminated (e.g., an incorrect response choice that more highly proficient subjects tend to select, but not as often as the key-correct response). The general approach to the item statistics in this phase is that some degree of weakness (small degree of relationship to total score) can be tolerated – the important thing is to establish the construct validity of the items as representatives of the ILR Skill Level Descriptions of proficiency.

The methods described in the following sections parallel those described by Schulz, Kolen, and Nicewander (1997, 1999). They present a general approach to using item response theory methods to develop number-correct scoring procedures for estimating level ratings from responses to specific sets of multiple-choice items (i.e., test forms). The fundamental concept that drives this development is that the levels represent a Guttman scale (Guttman, 1950). As Schulz, Kolen, and Nicewander express it: "Examinees at higher levels of achievement have mastered the same skills as those at lower levels, plus additional skills." The ILR levels of proficiency conform to this construct. The test development process, described next, is to verify that the item pool represents this construct, and to develop appropriate test forms from that pool.

[4] We have experimented with ways to assess English proficiency without greatly increasing the testing time, but have not yet found a satisfactory procedure.

Construct validity

Items for use in the multiple-choice tests have been developed with reference to the ILR Skill Level Descriptions of language proficiency. They are extensively reviewed by the development team, reviewers within the Test Development Division, as well as by a group of experts working for the Proficiency Standards Division (not part of the Test Development Division). The construct validity of the tests is demonstrated by showing that the items within one ILR level perform, on average, differently than the items at other ILR levels, across the range of proficiency. This is a multi-step process.

Estimating item parameters

In the first step, the items that are considered good candidates for inclusion in the calibration pool are identified based on the item analysis, and estimates of the item parameters of the three-parameter logistic model (Birnbaum, 1968) are calculated using the BILOG-MG program. The table below shows a sample of the output from this program.

Table 1 Sample of item parameter estimates

Item = item number
a = slope; b = difficulty; c = guessing; Level = ILR level
Status = K (Keep) or D (Drop) as representative of level
T = scored item in operational form
F = Not scored in operational form

Item	-a-	-b-	-c-	Level	Status	Form A	Form B
1	0.81484	-1.92824	0.25664	10	K	F	F
2	0.91512	-2.27842	0.25423	10	K	F	F
3	0.59444	-0.45490	0.24025	10	K	F	F
4	1.11142	-1.52041	0.25138	10	K	F	F
5	0.61634	-2.45859	0.27129	10	K	F	F
155	1.25545	0.82828	0.17826	30	K	T	F
156	0.52751	0.67925	0.33869	30	K	T	F
157	1.09816	0.26359	0.25530	30	K	F	F
158	0.86525	1.12015	0.30216	30	K	F	F
159	0.80151	1.20860	0.19200	30	K	F	F
160	0.00100	0.00000	0.25000	30	D	F	F
161	0.86758	1.88900	0.26902	30	K	F	F

The item parameters tell how the probability of answering correctly relates to proficiency. The more proficient the examinee, the more likely he/she is to answer correctly, but this relationship is not a straight line. Item response theory says the relationship is an "S-shaped" curve. At the low end of proficiency, gains in the probability of a correct answer are slow initially, but then accelerate, and then slow down once again as the probability of a correct answer approaches 1.00. The parameters are as follows:

a = item discrimination. This is the slope of the curve that relates proficiency to the likelihood of answering correctly; on some items (#5, above, for example) the slope is shallower than on others (#155). For the item with shallower slope, the probability of answering correctly does not increase with proficiency as rapidly as it would if the item had a steeper slope. Since the curve is not a straight line, this slope is defined to be the slope at the value of b, defined next.

b = the difficulty of the item. This is expressed as a logarithm of an odds ratio. Negative numbers mean that the item is rather easy, while positive numbers mean that the item is difficult. Graphically, b is the point at the middle of the S-curve where the rise stops accelerating and begins to decelerate (mathematically: the point of inflection).

c = the probability of guessing the correct answer if you have no proficiency. Item 155 seems to have a distractor that attracts low-proficiency examinees, so they are less likely to select the correct answer than if they were guessing at random.

Note that item 160 has a very shallow slope, zero difficulty level, and exactly the guessing probability one would expect for a 4-choice item. This item had a negative relationship with proficiency and the computer algorithm has declared that it is unrelated to proficiency. (Such items occur in nearly every test development project in any subject area; results like this cannot always be explained by examining the items.) This item was dropped – shown by the entry "D" in the Status column. Items marked K (Keep) in this column belong to the item pool from which the operational test forms will be drawn.

Given the three item parameters for an item, it is possible to calculate the probability that an examinee of any given level of proficiency will answer the item correctly. This is crucial to the next step of the process.[5]

Relating proficiency to expected percent correct

The column headed "Level" in Table 1 shows the ILR level of the item. All of the items to be included in the calibration pool are grouped by their nominal ILR level (assigned by the development team and verified in the review process). Within the first group, Level 1 (shown as 10 in Table 1, by convention), each item is taken in turn and the probability of answering that item correctly is computed across a range of proficiency levels. At this point, proficiency is an arbitrary value (called Theta, by convention) in a computational model. For our purposes, assuming a range of Theta from -4.0 through +4.0 spans a wider variation in actual proficiency than we are likely to see in practice (think of comparing four standard deviations below normal to four standard deviations above normal).

For each item (at this time considering only items at Level 1), we start at a proficiency of -4.0 and compute the probability of answering the item correctly. Since the set of items at level 1 are similar, but are not all exactly alike in terms of their item parameters (see Table 1), the probability that our hypothetical examinee with proficiency -4.0 will answer correctly will vary from item to item. The probability for each item will be less than 1, and for the examinee at the proficiency level of -4.0 they will be small because that person doesn't have much proficiency. If we add up all of these probabilities we have the expected score on this set of items for a person whose proficiency (theta) is -4.0. If we divide by the number of items in the set, we have the expected proportion correct for that person, which is 0.27 as shown in Figure 1. Next, we raise the proficiency level to -3.9 and re-compute the expected proportion correct at this new level on the same set of items. We continue incrementing the proficiency level in steps of +0.10 until we have computed the expected proportion correct on the items at ILR level 1 for examinees having proficiency of +4.0. The results for this series of computations are graphed in Figure 1.

Note that in Figure 1 proficiency is expressed on the Theta scale. The figure shows that a hypothetical examinee at proficiency level -4.0 (at the far left) is expected to answer only slightly

[5] Computations of cut-scores and scoring tables are accomplished by a special-purpose program developed by Dr. Daniel O. Segall of the Defense Manpower Data Center. This program also produces the output used to graph the relationship of expected percentage correct scores on proficiency, described next.

more than one item in four correctly – the individual with very little proficiency will not do much better than guessing randomly. As the hypothetical level of proficiency is raised, the expected proportion of items answered correctly increases. Somewhere around the proficiency level (Theta) of 1.00, the likelihood of a correct answer becomes very close to 1.00.

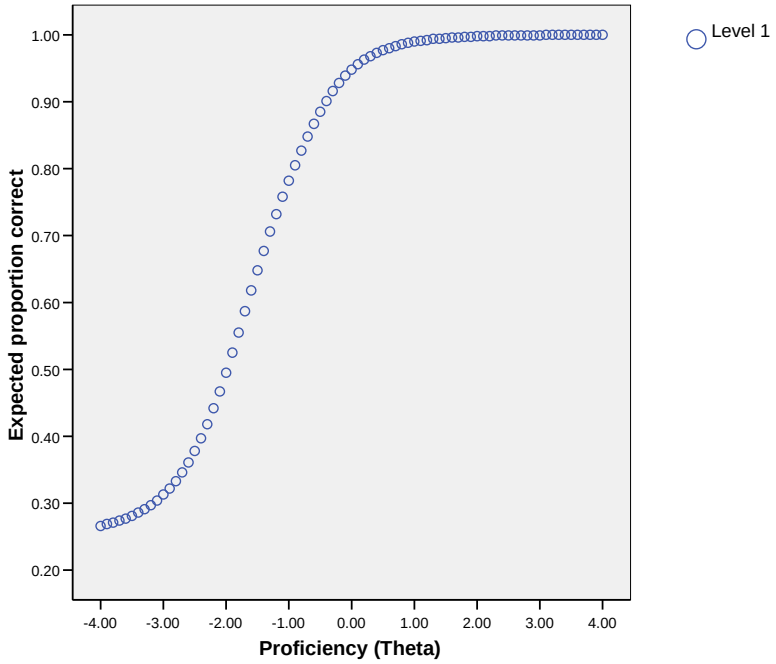

Figure 1 Expected proportion correct for item pool at ILR Level 1

The computations just discussed are then repeated for all of the items in the calibration pool at nominal ILR level 1+. We would expect, if the ILR Skill Level Descriptions of difficulty is correct, that the items at level 1+ would be more difficult to answer correctly and this should mean that the proportion correct for these items as a group will be lower than the proportion correct for item at nominal ILR level 1. Graphically, this means that the next set of dots should lie below and to the right of the set in Figure 1, except at the extremes: very proficient examinees will cope well with items at these levels, while examinees with very little proficiency will be guessing at the answers to questions at either level.

Graphical evidence of construct validity

In the next figures we add the curves for the items at subsequent ILR levels one by one. Each curve lies to the right of the previous curves – it requires more proficiency to attain the same percentage correct on the items at each level; that is, the items represent levels that are different from each other. Of course, for very proficient examinees (Theta of 3 or more), the curves all run together because those examinees are able to answer just about any question at ILR levels 1 through 3. Similarly, at the low end of proficiency, the curves bunch together because examinees at that level are guessing on all the items.

The fact that the curves for each successive ILR level do advance to the right demonstrates that the items in this calibration pool do have validity with respect to the construct of proficiency as defined by the ILR Skill Level Descriptions.

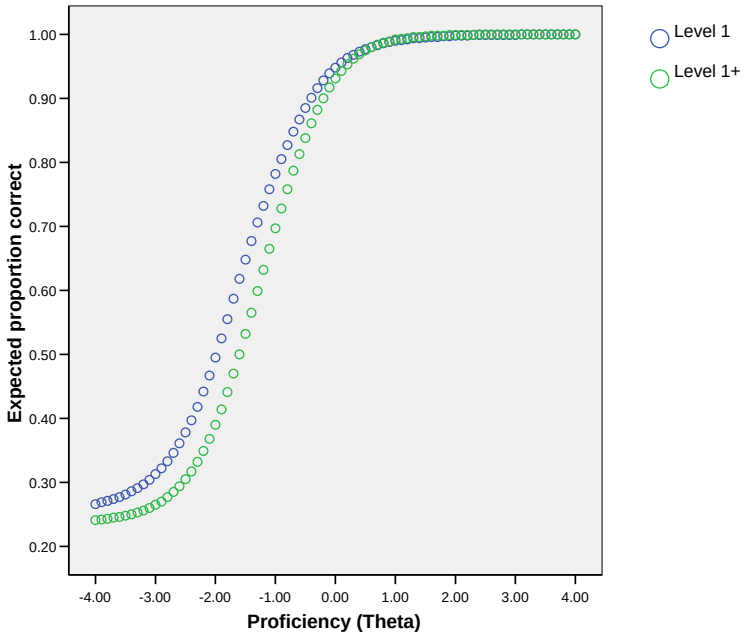

Figure 2 Item pools at ILR levels 1 and 1+

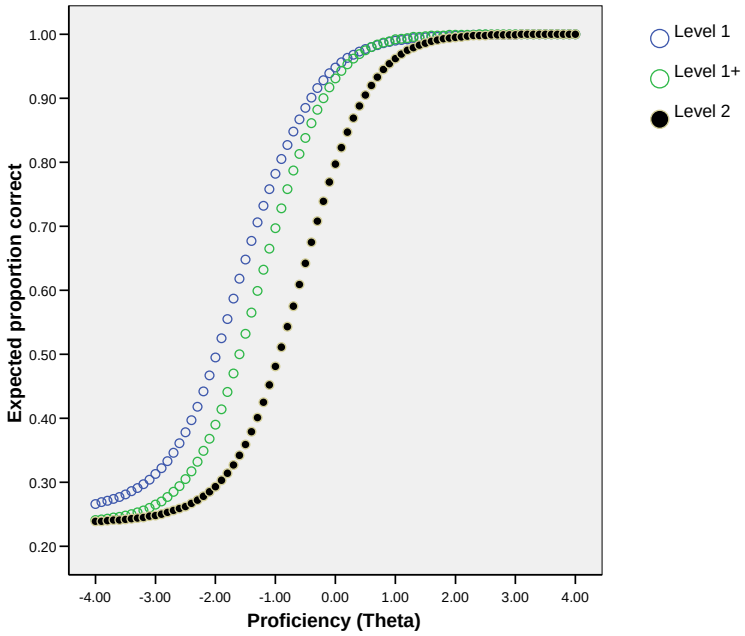

Figure 3 Item pools at ILR levels 1, 1+ and 2

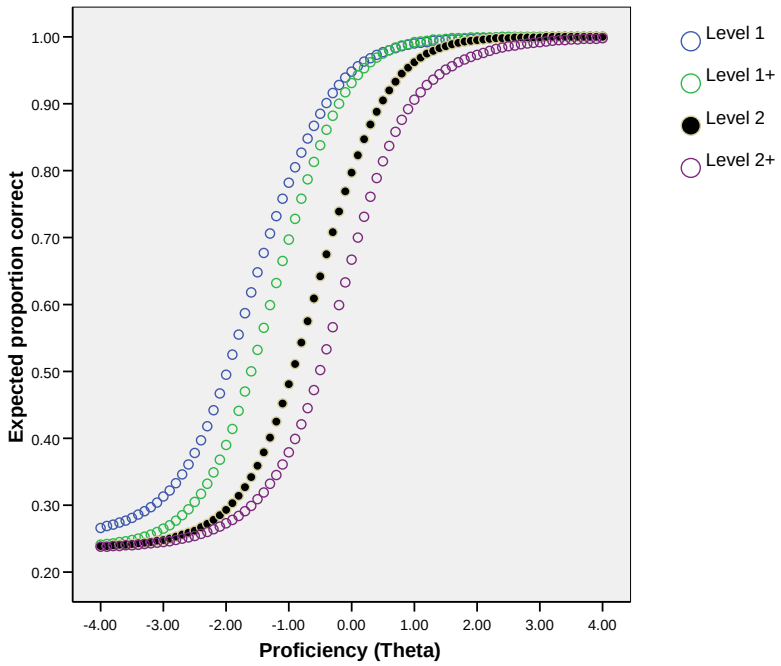

Figure 4 Item pools at ILR levels 1, 1+, 2, and 2+

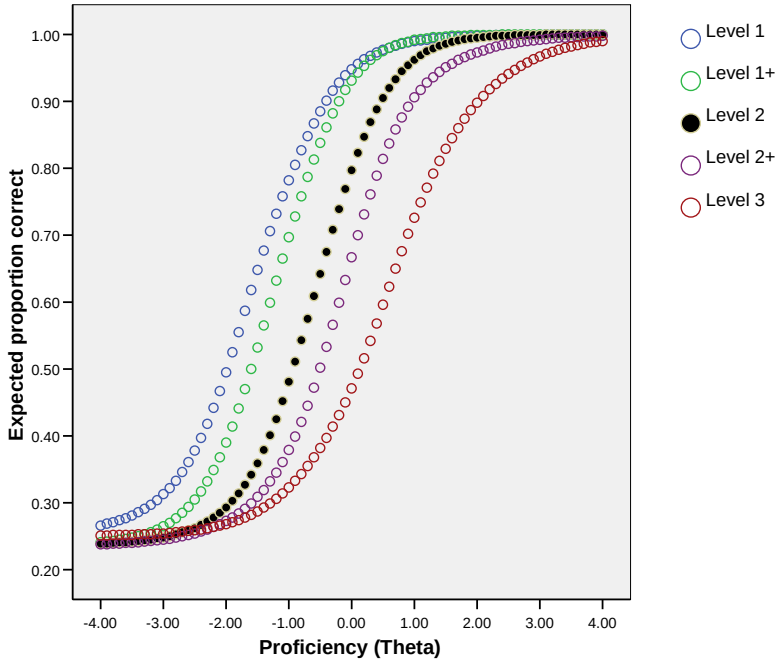

Figure 5 Item pools at ILR levels 1, 1+, 2, 2+, and 3

Determining proficiency cut-scores

Figure 5 is called the EPC-Theta Chart (which we pronounce as "epic-Theta") and shows the expected proportion correct on each of the item pools (defined by ILR level of the items) for the broad range of proficiency values (Theta) from -4.0 to +4.0. Through a process of consultation with experts in the ILR Skill Level Descriptions, it was determined that an examinee who has just crossed the threshold of proficiency into a particular level should be capable of answering 70 percent of the questions at that level correctly. In Figure 5 we can simply read from the chart the proficiency (Theta) required to attain 70 percent correct on the items at each level. We do this by drawing a horizontal line at 0.70 (expected proportion correct corresponding to 70 percent), and then noting the Theta value corresponding to the location where the 70 percent line crosses each of the EPC curves. This process is illustrated in Figure 6. The Theta cut score for Level 1 is -1.32.

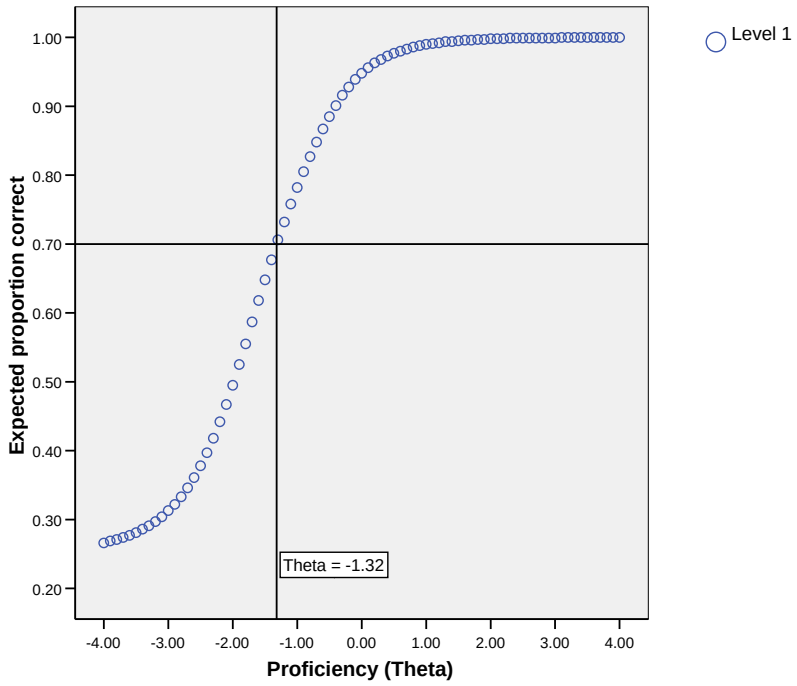

Figure 6 EPC-Theta curve for Level 1, with Level 1 Theta cut score

The method of successive approximations is used to compute the exact Theta cut scores (and these are shown in Table 2 and used in Figure 7). The values in Table 2 indicate the proficiency level required to reach the threshold of 70% correct on the group of items at each ILR level. Figure 7 "zooms-in" on the Theta values between -2.0 and +2.0 to show these lines more clearly.

Note that the Theta cut-scores increase with each threshold. This is the numerical confirmation of the construct validity of the item pools evident in Figures 5 and 7. In Figure 7 we can see that examinees at high levels of proficiency have mastery of lower levels – in fact, they are very proficient with lower-level material – confirming the definition of the Guttman scale given by Schulz, Kolen, and Nicewander (1997), quoted earlier. Appendix A illustrates this procedure in detail for the items at ILR level 1.

Table 2 Theta cut-scores based on the 70 percent mastery criterion

Cut-Scores based on 70 percent mastery criterion

	Theta
Cut-score between Levels 0+ and 1	-1.320
Cut-score between Levels 1 and 1+	-0.992
Cut-score between Levels 1+ and 2	-0.325
Cut-score between Levels 2 and 2+	0.101
Cut-score between Levels 2+ and 3	0.894

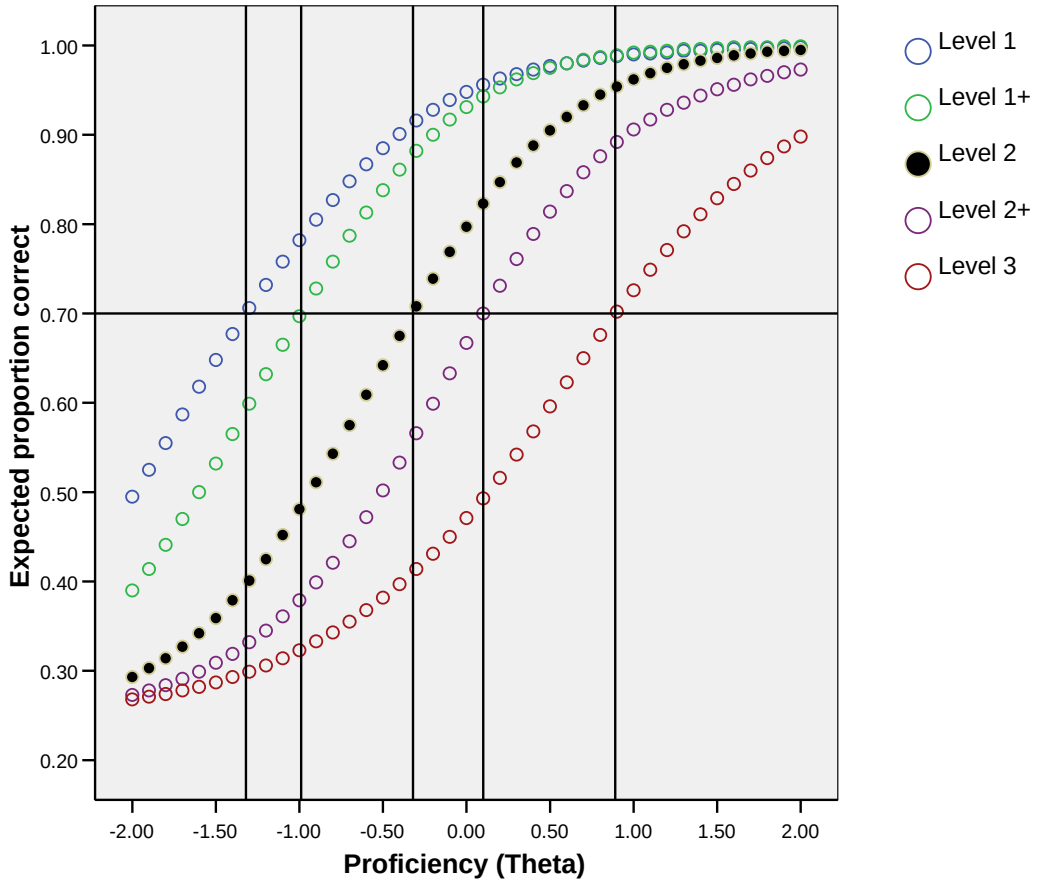

Figure 7 EPC curves for all ILR levels, with Theta cut scores

The next phase of development is to select items for the operational forms.

Item selection for operational forms

Each development team is supposed to produce two parallel forms in each tested skill (reading, listening) for operational use. The team leader is usually responsible for selecting the items and must simultaneously consider several guidelines. First, it is desirable that the overall level of each test be at approximately the same level of difficulty. This is accomplished by selecting very similar numbers of items from each ILR level for inclusion in each form. The test development specifications lay out the numbers of passages and items that should be included from each ILR level. The difficulty level of the items selected within each ILR should be approximately the same, on average, across the two forms.[6]

Second, the distribution of the content of the passages selected should be approximately the same in the two forms. This is usually accomplished by attending to the broad FLO (Final Learning Objective) content categories. All of these broad content areas should be included in both forms. The balance across content categories is constrained by the availability of items in the item pools (there may be only limited numbers of passages and items in some categories, for example).

Third, the items should be selected from among those that are the most discriminating as judged by the point-biserial correlation, or the magnitude of the "a" parameter in the logistic model.

Fourth, the probability of guessing the correct answer if you have no proficiency (the "c" parameter in Table 1) should be near 0.25 (we ask that items with values above 0.29 be avoided).

Within each passage, some items may have better difficulty and discrimination values than others. The team leader will have to balance these concerns in selecting passages for the operational forms.

Once the passages and items have been identified for the operational forms, the data collected for the calibration study can be processed once again to determine the cut-scores on the operational forms that will distinguish between the ILR levels.

Determination of number correct cut-scores

Before describing the process of determining the cut-scores in terms of number correct on the operational forms, we need to review the information we have for use in this process. First, we have the estimated parameters of the item response curves for the items we are going to keep in the calibration pool (see Table 1). Second, we have determined the Theta proficiency level required to be certified as mastering the items at each ILR level (Figure 7 and Table 2). What we need to do now is to determine the raw scores on each operational form that correspond to those Theta cut-scores.

Using the Theta cut score required to be certified as proficient at level 1, we can use the parameters of each item in Operational Form A to compute the probability of answering that

[6] This is not strictly necessary given the scoring methodology, but avoids creating the impression that one form is "harder" than the other. And, it serves to assure that the reliabilities of the two tests are similar throughout the score range.

item correctly. The probability of a correct answer to an easy item should be relatively high, while the probability of a correct answer to a more difficult item should be low. We then add up these probabilities to determine the raw score that would be attained by someone who is proficient at level 1. We can perform the same computation for the other levels, and repeat these computations for the items in Operational Form B. These values are shown in Table 3, which is an extension of Table 2. Appendix B details these computations for operational Form A.

Table 3 Number correct cut-scores for two operational test forms

Cut-Scores based on 70 percent mastery criterion

		-Number Correct-	
	Theta	Form A	Form B
Cut-score between Levels 0+ and 1	-1.320	17.808	17.058
Cut-score between Levels 1 and 1+	-0.992	20.833	20.457
Cut-score between Levels 1+ and 2	-0.325	29.645	29.893
Cut-score between Levels 2 and 2+	0.101	36.158	36.385
Cut-score between Levels 2+ and 3	0.894	45.661	45.266

Table 3 shows the proficiency (Theta) cut-score for each level (from Table 2) and the corresponding number correct cut score for each form. Theta can take on any value on the "real number line." But test raw scores are only given in whole numbers of items correct. So, we need a rule for converting the fractional cut-scores on the operational forms to the whole number scores that the tests will yield. An example will explain the process:

For Form A, the cut-score between 0+ and 1 is at 17.808. The raw score has to be higher than that to qualify for level 1, and the next higher raw score is 18. Thus, 17 becomes the highest raw score corresponding to level 0+, while 18 becomes the lowest score corresponding to level 1. The principle is to truncate the tabled raw score to find the highest raw score at the lower level, and then add one to the truncated score to find the lowest raw score at the higher level.

Table 4 Operational form scoring tables

Raw scores at each ILR level

Form A		Form B	
0	0 - 14	0	0 - 14
0+	15 - 17	0+	15 - 17
1	18 - 20	1	18 - 20
1+	21 - 29	1+	21 - 29
2	30 - 36	2	30 - 36
2+	37 - 45	2+	37 - 45
3	46 - 50	3	46 - 50

The cut-score between level 0 and level 0+ represents a special case. One way to calculate this value is to assume an examinee with very, very low proficiency (Theta of -5.0) and compute the score that such an examinee would attain. However, half of such examinees with essentially zero proficiency will be luckier than this, so we wanted to build in additional protection against classifying examinees with no proficiency as having 0+ proficiency. To do this, we use a computational procedure that computes the distribution of scores attained by examinees with zero proficiency (the procedure is described in Lord, 1980, p 44 and in Kolen and Brennan, 1995, p181), and we define the raw-score that only allows 20 percent of those examinees to attain level 0+ to be the cut score. Once the number correct cut-scores are determined, we can construct the scoring table (Table 4).

Assessing Reliability

Assessing internal consistency reliability for each form

The raw response data gathered from the calibration sample is processed again, this time using only the items in one of the operational forms. WINSTEPS computes an estimate of the KR-20 (also called the "Cronbach alpha") measure of the internal consistency reliability for this set of items. The procedure is performed on each of the two operational forms developed by the team. The results are compiled in (reference to document containing DLPT5 statistics, TBD).

Assessing parallel forms reliability

The proficiency level ratings provided each of the forms should agree to the point that one would feel comfortable using either form to obtain the rating. One form should not consistently produce higher ratings than the other form. Of greatest interest, for our purposes, is to be sure that the ratings agree with respect to which examinees should be assigned to level 1+ (or lower) or level 2 (or higher). Examinees must attain level 2 in order to graduate from the DLIFLC Basic class in any language, and must retain this level in order to earn proficiency pay (paid in addition to regular salary).

The ILR ratings obtained by applying the number-correct scoring rules to each of the operational forms are compared by arraying them in a cross-tabulation, as shown in Table 5.

Table 5 Cross-tabulation of ILR proficiency ratings from two operational test forms

Form A ILR Level Rating	Form B ILR Level Rating							
	0	0+	1	1+	2	2+	3	Total
0	10	2	0	0	0	0	0	12
0+	1	1	2	2	0	0	0	6
1	0	5	1	2	1	0	0	9
1+	0	2	1	9	6	0	0	18
2	0	0	0	5	18	3	1	27
2+	0	0	0	0	4	18	6	28
3	0	0	0	0	0	9	9	18
Total	11	10	4	18	29	30	16	118

Table 5 shows the ratings given to the 118 sample examinees using the two operational forms. Most of the values lie on the main diagonal (top left to lower right), indicating exact agreement. Most of the instances of disagreement are within a plus-level. When all examinees were classified as either 1+ and lower vs. 2 and higher on the two forms, 12 of the 118 examinees (10.2%) had different outcomes on the two forms. This is considered sufficiently accurate for our purposes.

A number of measures of parallel forms reliability may be derived from the data in this cross-tabulation. Table 6 shows the correlation statistics for this cross-tabulation. These statistics measure the degree of agreement between the two forms in different ways.

Table 6 Measures of correlation between ratings on parallel forms

Measure of Agreement		Value	Asymp. Std. Error (a)	Approx. T (b)	Approx. Sig.
Interval by Interval	Pearson's R	.918	.016	24.892	.000(c)
Ordinal by Ordinal	Spearman Correlation	.902	.018	22.522	.000(c)
Measure of Agreement	Kappa	.465	.054	11.262	.000
N = 118					

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.
c Based on normal approximation.

The Pearson correlation is the usual measure of agreement, and means that higher ratings on one form imply higher ratings on the other form. However, the two forms could be off by a constant value (one form could be consistently higher than the other), which would not be desirable, and would not be detected using this measure of agreement. The Pearson correlation is not responsive to our need to evaluate exact agreement.

The Spearman correlation indicates only the relationship of rankings on the two forms. Again, one form could be giving very different ratings than the other, but rank the values the same way (think of inches and centimeters, for example). This measure of agreement is also not responsive to our needs to evaluate exact agreement.

The Kappa value indicates the degree to which the forms give exactly the same rating under the assumption that there is no inherent ordering of the ratings. This is a kind of "worst-case-scenario." The previous slide showed that the discrepancies were usually very near the main diagonal, so the moderate value of Kappa is not very troubling.

Table 7 Intraclass correlation assessment of exact agreement

	Mean	Standard Deviation	N
Form A Level Rating	18.83	9.157	118
Form B Level Rating	18.88	9.049	118

Chi-Square Test of Goodness of Fit	Value	.126
	df	2
	Significance	.939

Under the strictly parallel model assumption

Intraclass Correlation Coefficient

	Intraclass Correlation (a)	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.918(b)	.885	.943	23.304	117.0	117	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.
a Type A intraclass correlation coefficients using an absolute agreement definition.
b The estimator is the same, whether the interaction effect is present or not.

Another method of assessing the degree of agreement is to use the intra-class correlation coefficient. This method allows us to examine whether the two forms meet certain statistical requirements for parallel forms and assesses the degree of exact agreement when those conditions are met. The relevant statistics are presented in Table 7.
The upper-left box in Table 7 shows that these two forms have about the same mean and standard deviation rating (on the ILR Skill Level Descriptions) across the 118 examinees who participated in the calibration. The upper-right box shows that the two forms passed a test indicating that they are strictly parallel forms – meaning that they are interchangeable: either one can be trusted to give almost exactly the same rating as the other.

A measure of the degree to which the two forms are interchangeable (meaning either one will give the same rating as the other) is the Intraclass Correlation Coefficient for Single Measures. The point estimate is 0.918 (on a scale that runs from 0 to 1.00). The confidence interval estimates a range that is highly likely to contain the true value. Based on the sample of participants in our study, there is a 95% likelihood that the true value for the Intraclass Correlation Coefficient lies between the lower bound (0.885) and upper bound (0.943). Our concern should be for the lower boundary, which should not decline very far below 0.85. These two forms lie within that boundary and have a satisfactory point estimate (0.918).

Summary

The calibration study conducted for each test development project provides the information needed to:

1. Select items for the calibration pools (one for each ILR level).
2. Determine whether the pools demonstrate validity with respect to the ILR construct.
3. Determine the proficiency level that is required to meet the ILR criterion for mastery.
4. Select items for the operational forms.
5. Determine the number correct cut-scores for each form and build scoring tables.
6. Assess the internal consistency reliability and parallel forms reliability of the forms.

Standards and guidance are provided for each of these steps and the results of these analyses are summarized for review and approval by authorized personnel.

References

BILOG-MG, Chapters 2 and 10 in M. du Toit, IRT from SSI, Lincolnwood, IL: Scientific Software International, 2003.

Birnbaum, A. Part 5 in Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores, Menlo Park, CA Addison-Wesley, 1968.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. A. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), Measurement and prediction (pp. 283-301). Princeton NJ: Princeton University Press.

Kolen, M. J. & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Linacre, J. M., A User's Guide to WINSTEPS, Chicago: MESA Press, 2002.

Lord, F. M. Applications of item response theory to practical testing problems. Hilsdale, NJ: Erlbaum, 1980.

Schulz, E. M., Kolen, M. J., and Nicewander, W. A. (1997) *A study of modified Guttman and IRT-based level scoring procedures for Work Keys assessments* (ACT Research Report 97-7). Iowa City, IA: ACT Inc.

Schulz, E. M., Kolen, M. J., and Nicewander, W. A. (1999) *A rationale for defining achievement levels using IRT-estimated domain scores*, Applied Psychological Measurement, *23*, 347-362.

Appendix A

Example computation of Theta Cut-Score

This appendix demonstrates one way to calculate the cut-scores that indicate the proficiency required on the Theta (latent proficiency scale) to attain 70 percent correct on a pool of items at a given ILR level. These values are subsequently used to compute the number correct scores on the operational forms that correspond to each ILR proficiency level. Appendix B demonstrates those computations.

There were 18 items at ILR level 1 included in the validation forms for this lower-range listening. One of them did not pass the validation review, leaving 17 usable items in the pool. The following table shows the input for the computation.

Table A-1 Parameters for level 1 items

Item	a	b	c
1	0.73658	-1.68834	0.27303
2	1.85174	-1.66029	0.25393
4	0.85385	-2.05297	0.25674
5	0.97997	-0.63746	0.27002
44	1.05538	-1.71196	0.22988
45	0.67949	-1.84538	0.25459
46	0.93694	-1.54670	0.24878
47	1.31941	-2.01004	0.24961
87	1.50381	-1.81942	0.23957
88	0.91906	-2.25533	0.26085
89	1.47097	-0.75069	0.20917
90	2.13943	-0.60882	0.30440
91	1.28287	-2.04145	0.24650
130	1.78043	-1.96742	0.23033
131	1.01509	-0.79753	0.28839
132	1.71717	-1.71414	0.21339
133	1.47800	-1.22187	0.25760

Each item is listed (identified by its number from the validation forms) along with the values for the three parameters estimated by BILOG-MG: "a" is the discrimination parameter, "b" is the difficulty parameter, and "c" is the estimated probability of answering correctly by chance alone.

The probability of answering correctly is determined by these three parameters and the proficiency of the individual taking the test, measured on the Theta scale. The formula is:

$$\Pr(\text{correct response}) = c + (1 - c)\,(e^{Q})/(1 + e^{Q}), \text{ where}$$

$$Q = 1.7a(\theta - b).$$

Table A-2 on the next page (extracted from an EXCEL spreadsheet) shows the probability of a correct response to each of these items when the value of Theta is set to -2.00. The sum of these probabilities is the expected number of correct responses to these 17 items. The expected proportion correct is computed by dividing 17 into the expected number of correct responses.

When Theta is set to -2.00, the expected proportion correct is considerably below 0.70 (equivalent to 70 percent correct). Table A-3 shows the probability of a correct response to each item in the level 1 pool when the value of Theta is set to -1.00. Now, the expected proportion correct is above the target value of 0.70. The values for the expected proportion correct in Tables A-2 and A-3 can also be found graphically using the curved line for level 1 in Figure 7 of the main document. The value of Theta corresponding to an expected proportion correct of 0.70 must lie between -2.00 and -1.00.

Table A-2 Computation of expected proportion correct when Theta = -2.00

Item	**a**	**b**	**c**	**Pr(Correct)**
1	0.73658	-1.68834	0.27303	0.566
2	1.85174	-1.66029	0.25393	0.445
4	0.85385	-2.05297	0.25674	0.643
5	0.97997	-0.63746	0.27002	0.338
44	1.05538	-1.71196	0.22988	0.518
45	0.67949	-1.84538	0.25459	0.594
46	0.93694	-1.54670	0.24878	0.494
47	1.31941	-2.01004	0.24961	0.629
87	1.50381	-1.81942	0.23957	0.534
88	0.91906	-2.25533	0.26085	0.703
89	1.47097	-0.75069	0.20917	0.242
90	2.13943	-0.60882	0.30440	0.309
91	1.28287	-2.04145	0.24650	0.640
130	1.78043	-1.96742	0.23033	0.596
131	1.01509	-0.79753	0.28839	0.368
132	1.71717	-1.71414	0.21339	0.451
133	1.47800	-1.22187	0.25760	0.350
			Expected Number Correct =	8.421
			Expected Proportion Correct =	0.495

Theta = -2.00

Table A-3 Computation of expected proportion correct when Theta = -1.00

Item	a	b	c	Pr(Correct)
1	0.73658	-1.68834	0.27303	0.784
2	1.85174	-1.66029	0.25393	0.917
4	0.85385	-2.05297	0.25674	0.868
5	0.97997	-0.63746	0.27002	0.528
44	1.05538	-1.71196	0.22988	0.832
45	0.67949	-1.84538	0.25459	0.796
46	0.93694	-1.54670	0.24878	0.778
47	1.31941	-2.01004	0.24961	0.929
87	1.50381	-1.81942	0.23957	0.917
88	0.91906	-2.25533	0.26085	0.909
89	1.47097	-0.75069	0.20917	0.485
90	2.13943	-0.60882	0.30440	0.440
91	1.28287	-2.04145	0.24650	0.930
130	1.78043	-1.96742	0.23033	0.961
131	1.01509	-0.79753	0.28839	0.583
132	1.71717	-1.71414	0.21339	0.913
133	1.47800	-1.22187	0.25760	0.730
			Expected Number Correct =	13.299
			Expected Proportion Correct =	0.782

Theta = -1.00

Using the "goal seeking" feature of EXCEL, we can ask that it find a value of Theta that results in the value of 0.70 for the expected proportion correct. The result of this computation is shown in Table A-4.

Table A-4 Computation (via goal seeking) of Theta corresponding to 0.70 expected proportion correct

Item	a	b	c	Pr(Correct)
1	0.73658	-1.68834	0.27303	0.719
2	1.85174	-1.66029	0.25393	0.809
4	0.85385	-2.05297	0.25674	0.809
5	0.97997	-0.63746	0.27002	0.447
44	1.05538	-1.71196	0.22988	0.745
45	0.67949	-1.84538	0.25459	0.737
46	0.93694	-1.54670	0.24878	0.691
47	1.31941	-2.01004	0.24961	0.868
87	1.50381	-1.81942	0.23957	0.834
88	0.91906	-2.25533	0.26085	0.861
89	1.47097	-0.75069	0.20917	0.362
90	2.13943	-0.60882	0.30440	0.353
91	1.28287	-2.04145	0.24650	0.870
130	1.78043	-1.96742	0.23033	0.905
131	1.01509	-0.79753	0.28839	0.494
132	1.71717	-1.71414	0.21339	0.811
133	1.47800	-1.22187	0.25760	0.583
			Expected Number Correct =	11.897
			Expected Proportion Correct =	0.700
	Theta =	**-1.320**		

This value can be confirmed in Figure 7 of the main document, and in Table 2.

This procedure may be applied to the pool of acceptable items at each level, to find the Theta cut-score defining the level of latent proficiency corresponding to the criterion level of success (70 percent correct for DLPT5 tests) at each level. Appendix B illustrates the procedure for using these Theta cut-scores to calculate the number of correct responses required on each operational form to attain each level of ILR proficiency.

Appendix B

Computing Number-Correct Thresholds on Operational Forms

Appendix A illustrated the procedure for computing the Theta cut-scores for each ILR level. Once those have been computed it is possible to evaluate outcomes on any combination of items from the total pool and determine the number correct corresponding to each ILR level. This procedure is illustrated in this appendix.

Table B-1 shows the items and parameters used in operational Form A of the lower-range listening test. In addition, the probability of correctly answering each item is shown for each of the Theta cut-scores listed in Table 2 of the main document. These probabilities are summed to give the expected number correct on this form for each level of the ILR. The procedure for converting these mixed numbers (whole number with decimal fraction) into the exact number-correct scores is described in the main document.

Table B-1 Computation of number correct scores corresponding to ILR levels

Item parameters				Theta 1	Theta 1+	Theta 2	Theta 2+	Theta 3
a	b	c	ILR Level	-1.320	-0.992	-0.325	0.101	0.894
0.73658	-1.68834	0.27303	10	0.719	0.786	0.888	0.930	0.972
1.50381	-1.81942	0.23957	10	0.834	0.918	0.984	0.994	0.999
2.13943	-0.60882	0.30440	10	0.353	0.443	0.817	0.951	0.997
1.01509	-0.79753	0.28839	10	0.494	0.585	0.782	0.875	0.964
1.71717	-1.71414	0.21339	10	0.811	0.915	0.987	0.996	1.000
1.45680	-0.56819	0.19330	16	0.302	0.402	0.715	0.871	0.979
1.42974	-1.26499	0.19044	16	0.568	0.725	0.925	0.972	0.996
1.39705	-1.38363	0.24134	16	0.649	0.785	0.943	0.978	0.997
1.08915	-0.87757	0.22496	16	0.462	0.572	0.795	0.891	0.972
1.45077	-0.45848	0.26233	16	0.341	0.418	0.691	0.852	0.975
1.63135	-0.28344	0.24402	20	0.284	0.337	0.600	0.806	0.972
1.85018	-0.15339	0.32695	20	0.344	0.372	0.575	0.791	0.976
0.97588	-0.52548	0.25428	20	0.412	0.490	0.689	0.805	0.935
2.02921	-0.26799	0.17893	20	0.200	0.241	0.549	0.820	0.985
1.35521	-0.70748	0.23991	20	0.389	0.500	0.777	0.898	0.981
1.83511	-0.61061	0.23144	20	0.307	0.411	0.776	0.925	0.993
1.93750	-0.53776	0.30368	20	0.353	0.431	0.769	0.924	0.994
1.81910	0.31692	0.32716	20	0.331	0.339	0.408	0.555	0.903
1.30874	0.78520	0.24311	20	0.250	0.257	0.302	0.379	0.667
1.18396	-1.13399	0.25246	20	0.557	0.679	0.877	0.943	0.988
2.22328	-0.10089	0.14881	20	0.157	0.177	0.404	0.729	0.981
1.48781	-0.86416	0.22304	20	0.409	0.549	0.842	0.938	0.991
2.55103	-0.50669	0.20724	20	0.230	0.293	0.752	0.947	0.998
1.51959	-0.55397	0.19598	20	0.294	0.392	0.714	0.875	0.981
2.19742	-1.25060	0.24433	20	0.573	0.792	0.977	0.995	1.000
1.76289	-1.10770	0.24223	20	0.504	0.686	0.934	0.980	0.998
1.81177	0.00034	0.29334	26	0.305	0.325	0.483	0.701	0.958
0.85810	-0.16975	0.26454	26	0.380	0.435	0.591	0.704	0.871
1.52725	0.19272	0.19537	26	0.211	0.231	0.362	0.550	0.888
1.50570	0.11693	0.22188	26	0.241	0.265	0.412	0.603	0.906
1.42026	-0.21178	0.22935	26	0.279	0.331	0.562	0.754	0.950
1.72796	0.50160	0.14217	26	0.146	0.153	0.212	0.344	0.794
1.34358	0.36265	0.27265	26	0.288	0.304	0.398	0.531	0.833
2.27929	0.49960	0.14626	26	0.147	0.149	0.180	0.296	0.848
2.18245	0.44708	0.22248	26	0.224	0.226	0.264	0.391	0.876
1.97811	0.02148	0.18733	26	0.196	0.213	0.381	0.648	0.959
1.51045	0.10627	0.27295	26	0.291	0.314	0.454	0.634	0.915
1.89880	-0.26995	0.22940	26	0.255	0.298	0.581	0.821	0.982
1.11582	-0.89511	0.28088	30	0.503	0.607	0.818	0.906	0.977
3.59773	0.37110	0.22604	30	0.226	0.226	0.237	0.351	0.970
1.64661	0.20570	0.27960	30	0.290	0.304	0.413	0.587	0.908
0.78072	0.03724	0.25287	30	0.359	0.405	0.538	0.642	0.819
1.19545	0.83531	0.19648	30	0.206	0.216	0.266	0.344	0.622
1.53720	0.54993	0.30269	30	0.308	0.315	0.367	0.467	0.798
1.26949	0.39358	0.25543	30	0.273	0.291	0.386	0.514	0.811
0.90466	0.33142	0.26575	30	0.319	0.351	0.462	0.568	0.782
0.78488	-0.01923	0.22275	30	0.339	0.389	0.533	0.642	0.823
0.76906	1.42435	0.36725	30	0.384	0.393	0.426	0.463	0.578
1.43392	0.65738	0.14950	30	0.156	0.164	0.221	0.324	0.694
0.91464	-0.40267	0.19943	30	0.354	0.428	0.624	0.749	0.906

Expected number correct:	17.811	20.829	29.641	36.156	45.663
First raw score at level:	**18**	**21**	**30**	**37**	**46**
Last raw score at level:	**20**	**29**	**36**	**45**	**50**

Appendix C: Lower-Range DLPT5 Constructed Response Test Scoring Procedures

Lower-Range DLPT5 Constructed Response Test Scoring Procedures

Summary

1. Examinees enter answers on the computer.
2. Examinee answer files and identifying information are sent to a central scoring manager.
3. The scoring manager assigns two independent scorers to score the test and gives each of them in turn the examinee response file (printed or electronic), a scoring sheet, and the scoring protocol.
4. Each scorer evaluates the answers according to the protocol and enters the appropriate value ("1" for a correct answer and "0" for an incorrect answer) on the scoring sheet.
5. Each scorer calculates the examinee's level score according to the scoring formula and enters the level score on the scoring sheet.
6. All materials are then returned to the scoring manager.
7. The scoring manager checks to make sure the two levels generated by the two scorers are the same, and to make sure the level scores were computed correctly from the number of correct responses.
8. If there are discrepancies or errors, the scoring manager assigns a third rater, who scores the test again without referring to either of the previous scorers' work. The third rater should be a "senior scorer," someone whose judgment is highly reliable.
9. The scoring manager examines the third rater's scoring sheet, and the resulting score is awarded to the examinee. In rare cases, the third rater's score may differ from that of both other scorers; in such a case, a fourth rating is necessary.
10. Item analysis data from the scoring sheets are sent to the ES psychometrician for analysis.