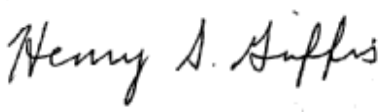
Readiness Metrics in Support of the Defense Language Program

Peggy A. Golfin • Jessica S. Wolfanger • Peter H. Stoloff
with
James E. Grefer • Darlene E. Stafford

DRM-2012-U-001244-Final
September 2012

CNA
ANALYSIS & SOLUTIONS

Approved for distribution: September 2012



Henry Griffis, Director
Defense Workforce Analyses
Resource Analysis Division

This document represents the best opinion of CNA at the time of issue.
It does not necessarily represent the opinion of the Department of the Navy.

Approved for Public Release; Distribution Unlimited. Specific authority: N00014-11-D-0323
Copies of this document can be obtained through the Defense Technical Information Center at www.dtic.mil
or contact CNA Document Control and Distribution Section at 703-824-2123.

Contents

Executive summary

The mission of the Defense Language and National Security Education Office (DLNSEO) is to provide strategic guidance on present and future requirements related to language, regional expertise, and culture (LREC) [1]. DLNSEO's duties include tracking and reporting on the accession, promotion, retention, and attrition of personnel with language skills, of language professionals, and of Foreign Area Officers (FAOs). The office asked for our help in developing these and other metrics in support of its mission, and in support of achieving the goal of having the required combination of LREC capabilities to meet current and projected needs.

Fill and FIT

Based on our review of the literature concerning effective metrics (especially those relating to readiness) and our interviews with various LREC stakeholders, we developed two metrics that we consider to be the foundation for determining the status of language readiness:[1] a measure of the current number of servicemembers with any level of proficiency that could fill current and contingency requirements (fill) and a measure of the extent to which these servicemembers satisfy the full range of these requirements in terms of proficiency in all language modalities, paygrade, service, and so on (FIT).[2]

When these two metrics are calculated quarterly for each language,[3] and within each by language modality, service, and other important

1. Language proficiency is currently the only one of these skills that is measured and documented in personnel records. Our metrics are applicable to the other capabilities.
2. By convention, "fill" is lower case, while "FIT" is all caps.
3. Language proficiency is differentiated by digraphs, which are in transition in various databases to trigraphs. For simplicity, we use the word *language* to refer to digraphs and trigraphs.

characteristics of requirements, they provide strategic guidance as to whether capabilities are too low and, if so, where the deficiencies are, in terms of these characteristics.

Pinpointing causes and early warning metrics

We proposed additional metrics that satisfy two important properties of good metrics: the ability to (1) drill down to more detailed information and pinpoint causes of problems that are identified in fill and FIT calculations (such as in recruiting, training backlogs, and attrition) and (2) provide early warning that deficiencies in LREC capabilities might arise in the near term to the longer term. These metrics are also important in determining whether the Total Force is trending toward an LREC "hollow force," which is a specific concern expressed to us by Department of Defense leadership.

Aggregation

There are a few other properties of good metrics that we used as guidance. In particular, we recommended against metrics that result in misleading aggregation. For instance, reporting the number of servicemembers with *any* proficiency (including proficiency that is only self-professed but never formally tested) in *any* language (including languages that are not of strategic importance) provides very little useful information. The goal of the Defense Language Program is to have enough of the *right* people with the *right* skills, not simply to have as many servicemembers as possible with any level of skill in any foreign language.

We also caution DLNSEO against using metrics that are based on too aggregated a population. For instance, measuring the retention of proficient servicemembers in isolation, without comparing their retention to their peers, is misleading and ignores the reality that many of the services are in the process of downsizing. The better metric is whether more proficient servicemembers are leaving relative to their otherwise similar peers, and of special concern is the relative loss of those with proficiency in languages of the greatest strategic importance.

Goals

We intentionally avoid establishing goals for these metrics. Until the language requirement process is complete, it is not possible to know whether there are too few, too many, or the right number of servicemembers with language proficiency. The setting of goals is also beyond any current understanding of the effect of deficits of language or culture capabilities on readiness. For example, little is known of the consequences of having too few speakers of a language at Interagency Language Roundtable (ILR) level 2 on the ability of a particular unit to perform its duties. Goals need to be established on the basis of the acceptable level of risk that leadership is willing to assume if requirements fall short by, say, 5 or 10 percent. Research necessary to establish such goals is lacking and generally not possible because the data to conduct the analysis either do not exist or are not readily available. For instance, the benefits of increased LREC capabilities for General Purpose Forces may be outweighed by the readiness and/or financial costs of attaining that level of LREC skills. The time servicemembers spend in LREC training is time spent away from full duty and from performing and enhancing their primary occupation skills. We conclude that research of the costs and benefits of additional LREC capability is very important, but LREC goals should not be established until that research has been conducted.

Internal data collection and analysis

We propose that DLNSEO obtain and manage the data necessary to generate the metrics we propose and others that are required by directive. DLNSEO currently relies on services and other entities to provide inputs for these metrics, but there is little consistency across the services and over time in how many metrics are measured and reported. Some of the important properties of metrics that we noted are that they are consistent and reliable, and allow comparisons across organizations and over time. With so many different methods currently in use, none of these properties are possible. Thus, it is also not possible to satisfy some of the other properties of good metrics—that they are useful in charting progress toward goals and are useful in evaluating the impact of innovations or changes in policies.

A dedicated "LREC skill manager," similar to the service's occupation managers, should be appointed to perform these duties, which would include obtaining all of the relevant data from the Defense Manpower Data Center, the services, Defense Language Institute's Foreign Language Center (DLIFLC), and so on, and to produce the required reports. We submit that only by becoming intimately familiar with these data will DLNSEO be able to provide the full range of support and oversight required by the Defense Language Program.

Finally, we propose that DLNSEO produce and disseminate a quarterly report that summarizes the fill, FIT, and early warning metrics that we recommend. Some policy interventions that may be required to address problems are under the purview of DLNSEO, such as the Foreign Language Proficiency Bonus (FLPB), but many are not, including recruiting and retention goals, enlistment and retention incentives, and training. We submit that the role that DLNSEO serves in this capacity is in the timely dissemination of early warning of problems to the appropriate leaders so that they can address the problems that are under their purview.

Background

U.S. military operations over the past decade have highlighted the importance of ensuring that our military personnel have the right foreign language, regional expertise, and cultural (LREC) capabilities to meet current and emerging requirements. The Department of Defense (DOD) has undertaken substantial efforts to make certain that our military has sufficient organic LREC capabilities to ensure our nation's security.

In support of these efforts, DOD published the Defense Language Transformation Roadmap (DLTR) in 2005 to "provide to the Deputy Secretary of Defense...a comprehensive roadmap for achieving the full range of language capabilities necessary to support the 2004 Defense Strategy" [2]. The roadmap called for (a) establishing metrics to monitor performance of the Defense Language Program (DLP), including metrics on the use and management of language skills and on the accession, promotion, and retention of personnel with language proficiency, and (b) instituting a process for regular reporting to the USD (P&R).

The DLTR also called for the establishment of a language office within the Under Secretary of Defense for Personnel and Readiness (USD (P&R)) to

> ensure a strategic focus on meeting present and future requirements for language and regional expertise. This office will establish and oversee policy regarding the development, maintenance, and utilization of language capabilities; monitor trends in the promotion, accession and retention of individuals with these critical skills; and explore innovative concepts to expand capabilities.

This office, the Defense Language Office (DLO), was established in 2005.

In 2012, DLO merged with the National Security Education Program (NSEP) to become the Defense Language and National Security Education Office (DLNSEO). DLNSEO asked CNA to help with establishing metrics specified in the DLTR and other documents.

In the first phase of this study, we researched the roles and responsibilities of DLNSEO and reviewed existing reports and data in support of the DLTR. We conducted a literature review to identify what makes a good set of metrics, both in general and specifically for readiness reporting, and we interviewed stakeholders to identify what metrics they viewed as important for tracking the progress of the program. We then developed a number of metrics that are based on data that are available currently or that we recommend should be obtained in order to satisfy several of DLNSEO's roles and responsibilities.

While DLNSEO's oversight includes language, regional expertise, and cultural proficiency, language proficiency is the only one of these skills that is currently measured and documented in personnel records. As a consequence, we focus our research on language metrics specifically, but our approach and recommendations are generally applicable to the remaining skills once they are documented.

The paper is organized as follows. We begin our discussion with an overview of some of the most important components of DLNSEO's current language metrics and the tools used to derive them. We then summarize our findings from our review of the literature and stakeholder interviews, which establishes essential features of DLNSEO metrics that help to guide our work. Next, we turn to a discussion of the metrics we propose and how they can be coordinated with other current DLNSEO language reporting efforts. We conclude with recommendations for DLNSEO regarding who should manage the data and produce the reports we propose.

DLNSEO overview

To develop metrics for DLNSEO, we need to understand the scope of its authority and mission. The following two excerpts from the DLNSEO website [1] state its vision and mission, respectively:

> The Department will have the required combination of language, regional, and cultural capabilities to meet its current and projected needs.
>
> Provide strategic direction and programmatic oversight to the Military Departments, Defense field activities and the Combatant Commands on present and future requirements related to language, regional expertise, and culture.

Data and reporting requirements

A number of organizations share some or all of the DLNSEO's vision, and each has specific oversight and requirements. For the purpose of creating metrics that support DLNSEO's vision and mission, we refer to instructions and directives listed on the policy portion of its website that state specific metric requirements [3, 4, 5]. In summary, DLNSEO is required to do the following:

1. Develop measures that evaluate progress toward the goal of increased language and regional proficiency capabilities throughout the department.
2. Provide guidance for foreign language incentives.
3. Track the accession, promotion, retention, and attrition of personnel with language skills of strategic interest to the department.
4. Develop and sustain personnel systems that maintain accurate data on all DOD personnel with certified and self-reported foreign language proficiency and area expertise.

5. Determine when there is a "critical need."
6. Publish a DOD strategic language list and update it as required.
7. Establish a language readiness-reporting index to measure language capabilities within the DOD components.
8. Monitor the accession, retention, and promotion of language professionals.
9. Establish metrics and monitor FAO accession, retention, and promotion rates.

In early 2011, DOD published a plan that provided strategic guidance for how the Total Force could expand LREC capabilities and improve the effectiveness of servicemembers with those skills through 2016 [6]. The plan specified three goals that represented the top LREC priorities: (1) establish LREC requirements, (2) build and sustain a Total Force with the right LREC capabilities to meet existing and emerging requirements, and (3) strengthen LREC skills to increase interoperability and build partner capacity. Because the second goal is the one most aligned with the types of metrics for which DLNSEO is responsible, it is the one we focus on in the present study.

Overview of current language data and reports

Because many of the DLNSEO metrics requirements have been in existence for several years, they have produced a number of reports. Our work is intended to help the office refine some of these and to provide additional metrics and data. It is necessary, therefore, to describe some of the data and reports already in use by DLNSEO. We begin with a discussion of the language readiness reporting system required in [4].

Defense Readiness and Reporting System and Language Readiness Index

The ability to manage readiness has become increasingly important over the last decade as DOD has faced the challenges of finding enough units that were adequately trained to fulfill both steady-state and emerging requirements. These challenges led DOD to revise the

way it had traditionally thought about and measured readiness, resulting in a transition from a readiness system based on resources to one based on capabilities,[4] and one that focuses more on the implications of deficiencies than on the deficiencies themselves [7].

In 2002, DOD created the Defense Readiness Reporting System (DRRS) in conjunction with this new approach to readiness. The DRRS is an internet application that provides a capabilities-based requirements reporting system that, according to the DRRS website, allows users to "evaluate the readiness and capability of U.S. Armed Forces to carry out assigned tasks. That is, to find units that are both ready and available for deployment in support of a given mission" [8].

DRRS includes a special component devoted specifically to LREC readiness—the Language Readiness Index (LRI). The LRI satisfies both the requirements in [4], which specifies that a database must be created to track language skills and capabilities, and the requirements specified in the 2006 Chairman of the Joint Chiefs of Staff Instruction (CJCSI) 3126.01 [9]. One of the DLNSEO mandates in this instruction is the monitoring of LREC requirements:

> DLO will consolidate all COCOM language requirements into one database that will represent a quarterly "snapshot" of reported language needs. The DLO will develop a secure Web-based capability to collect and organize the data provided by the COCOM and establish a process to provide the Joint Staff, COCOMs, Services, and Defense agency representatives access to this information. The DLO will use this data as the basis for forming language and regional expertise policy guidance. [9]

LREC requirements are undergoing major revision, as part of a Capabilities Based Requirements Identification Process (CBRIP) that began a few years ago. This process, which is being led by the Joint Staff in conjunction with the combatant commands, has developed a way to identify language and regional expertise capability requirements as well as a process to send a demand signal to the services.

---

4. In this regard, a capability is the ability to perform a given task to specified standards by either a parent organization or by operational needs.

The Joint Staff recently completed the first phase of this process, focusing on the geographic combatant commands' steady-state requirements, and the second phase, identifying the geographic combatant commands' surge requirements. The third phase, set to conclude in FY 2015, will deal with all remaining requirements, to include those of the functional combatant commands and contingency operations. Because the entirety of LREC requirements will undergo significant changes in the next few years, and current requirements are considered to be fairly incomplete, LRI is best used currently as a method to determine LREC capabilities rather than as a tool to determine LREC readiness.

Ultimately, however, once the CBRIP is complete, according to its website, LRI will serve primarily to identify gaps in language readiness resource needs. We refer interested readers to the LRI website for more details regarding the specific types of information available now, and proposed for the future. We will refer to some of the details of various components later, as they relate to our metric recommendations.

Language reports

One of the most comprehensive relevant reports that DLNSEO produces is in fulfillment of DODI 5160.70. This annual report, which provides metrics regarding the accession, promotion, retention, and attrition of personnel with language skills in the department, is based on data collected from the services, defense agencies, Defense Manpower Data Center (DMDC), and other sources. The latest report available to us was the 2010 Annual Foreign Language Report [10].

This report notes that the data calls from the services/agencies that are required by the instruction

> continue to be inconsistent with the Department's DMDC data. This inconsistency continues to create challenges in determining a valid baseline from which the department can effectively assess and evaluate impacts of programmatic and/or policy changes to the various language programs. [10]

The report also states the following:

> Reporting by the Services remained restricted to limited military occupational specialties and/or programs. This reporting restriction is not in compliance with DODI 5160.70 reporting guidance since the requirement is for the Services to report on all personnel who have a language capability. This lack of compliance means that those personnel with a language capability who do not possess one of the select military occupational specialties are not counted. [10]

We will refer to this report later, as we offer recommendations for modifications to some of the data used, and metrics constructed, in order to provide more useful reports.

How our work enhances DLNSEO's efforts

DLNSEO has made significant progress toward satisfying the data and reporting requirements specified in various instructions and directives, in a relatively short period of time. Most of these efforts are evolving, as DLNSEO incorporates lessons learned and awaits better data from the services and others.

In support of these efforts, we focus our work on two things:

1. Developing metrics for DLNSEO that (a) track accession, retention, and promotion rates of enlisted language professionals (LPs),[5] FAOs, and all other servicemembers with LREC proficiency, (b) track language readiness, and (c) are helpful in revising the Strategic Language List (SLL) and determining when there is a "critical need"
2. Advising DLNSEO regarding the necessary data to derive and maintain these metrics

We turn next to our literature review of the properties of good metrics and results of our interviews with stakeholders.

5. We define *language professionals* as belonging to the enlisted ranks only. FAOs are not language professionals per se. Henceforth, when we refer to LPs, we mean enlisted LPs only.

This page intentionally left blank.

What makes a good metric

Part of our tasking was to conduct a review of the literature to determine the desirable properties of a good metric and to interview stakeholders in order to understand their concerns and what they considered to be important properties of the metrics we develop.

We begin by defining what we mean by a *metric:* it is a standard definition of a measurable quantity that indicates some aspect of performance. For instance, one language metric would be the number of servicemembers who have any proficiency[6] in a particular language.

The term *performance metric* is sometimes used to refer to measuring progress toward a performance objective or goal. In the case of DOD language capabilities, we conclude that this is the right approach to establishing metrics, using DLNSEO's vision and mission as specific objectives. More specifically, our proposed metrics will provide a measure of, and tools to manage, language readiness.

We turn next to our literature review of what makes a good metric.

General properties of good metrics

Our literature search revealed some common attributes of good metrics, regardless of the industry or mission of the organization, that we believe DLNSEO metrics should possess [11 through 22]. Specifically, a good metric should:

- Be useful in charting progress toward ultimate objectives (i.e., the DLNSEO's vision and mission)

6. We assume that the reader is familiar with how language proficiency is measured and reported by DOD, based on the Interagency Language Roundtable (ILR) guidelines. We refer those who are not familiar to the ILR website: www.govtilr.org.

- Not produce unintended consequences[7]
- Provide early warning signals of potential problems in a timely manner (e.g., language training backlogs, increased losses of language proficient servicemembers, or decreased testing)
- Be useful in evaluating the impact of process innovations and changes in performance, such as new training methods and changes in FLPB
- Be consistent and reliable across all levels of the organization to allow comparisons across organizations and time
- Not cost more to generate, in terms of the amount and difficulty of data necessary, than the benefit they provide

We turn now to a review of some of the most relevant readiness metric research that provides additional guidance for our proposed metrics, much of which was conducted in support of the development of the DRRS.

Review of readiness metrics

Production function approach

In order to construct metrics for the DRRS, the authors of [23] described readiness as something similar to the production element of a firm. Because of this, they chose readiness metrics that "cover the entirety of the production function by measuring inputs in terms of their contribution to outputs at each stage of the process." They argued that this approach allows for the tracking of readiness status over time, identification of important variations, and appropriate diagnoses of problems. Their conclusions are consistent with business metrics literature that we reviewed previously; metrics should be designed in a way to allow managers to pinpoint causes of problems.

7. For example, measuring only the number of heritage speakers recruited in a service may cause recruiters to disproportionately focus their efforts on these types of recruits, which could result in a failure to meet goals for other types of recruits.

They also concluded that metrics should be developed that limit the number of signals given to top leadership; this group should be provided with the most important metrics to allow for appropriate action, including those metrics that provide an early warning of decreasing capabilities. More detailed metrics, especially ones that cover the entire production process in more detail, are more relevant for lower levels of the organization.

Applying their production function approach to language readiness metrics means that all of the additions and losses to language readiness are measured and monitored. Additions to the pool of language-proficient servicemembers are derived from a number of sources, such as new accessions with language proficiency, those who receive language training, and the number revealing their proficiency by testing. Losses are derived not just from proficient servicemembers leaving the Total Force but from degradation of proficiency, lapsed testing, changes in the number available for assignment, and so on.

The advantage of this approach is that it allows for the precise identification of potential problems and the tracking and monitoring of both the current status of language capabilities and projected future inventories. The ability to project future inventories is vital for the establishment of early warning signals of potential language readiness degradation in the near term and longer term because language requirements can change rapidly and the time to train many foreign languages is often very lengthy.

Aggregation

Other research on readiness metrics highlights the dangers of simply aggregating metrics into one summary metric. For instance, the authors of [24] argued that the tendency to aggregate readiness metrics into a relatively few or even a singular metric is often misleading, especially if the arithmetic mean is used to calculate the aggregated metric since it assumes that inputs are substitutes. They provided a compelling example of this problem: Assume that the battlegroup is ready in every aspect except that there are no F-14s/FA18s; the arithmetic mean would show that mission readiness was reduced by only 2.4 percent, down to 97.6 percent.

To be more meaningful, they recommended that (1) aggregation should not cross mission areas, (2) weights should be applied to components to avoid the example they cite, (3) aggregation does not necessarily have to produce just a single number, and (4) the arithmetic mean is a flawed aggregation tool. We will return to this notion of alternatives to the arithmetic mean in calculating readiness later, when we present some of our proposed language readiness metrics.

The caution against aggregating metrics in [24] is especially important in considering language readiness metrics. Not all languages have the same number of requirements, and some languages for which testing is available have no requirements. Reporting an aggregated metric of the number of proficient servicemembers in all languages—even those for which there are no requirements and those for which the supply of proficient servicemembers far exceeds requirements—is at best an uninformative metric and at worst a misleading metric that could have serious readiness implications because it masks serious deficiencies in languages of strategic importance.

As an example, consider the case in which the services either survey all servicemembers who identify themselves as Hispanic to determine whether they have any familiarity with Spanish or require each of these servicemembers to take a Defense Language Proficiency Test (DLPT) in Spanish. Spanish is an Enduring language according to the SLL, and there are a sufficient number of servicemembers proficient in Spanish to fulfill requirements.[8] Either action would greatly increase this aggregated metric of servicemembers who are proficient in *any* language, even if there was no change in the number of servicemembers proficient in all other languages. The increase could be viewed by leadership to indicate that language readiness has improved or that more servicemembers are learning a foreign language, perhaps because of FLPB or other policies. Neither conclusion would be correct, however.

8. The SLL categorizes languages as (1) Immediate (immediately needed to meet urgent demands), (2) Emerging (anticipated expanding future requirements,), or (3) Enduring (a continuing need for the next 10 to 15 years) [25]. In general, DOD is lacking a sufficient number of servicemembers who are proficient in languages in the first two categories.

A more serious and misleading conclusion would arise if the increase were accompanied by a simultaneous but somewhat smaller *decrease* in the number of servicemembers who have a tested proficiency in languages on the Immediate or Emerging list of the SLL. The aggregate metric would have increased, but readiness would have actually decreased because there would be fewer proficient members in these languages to fill requirements.

The bullets below summarize all of the properties of good metrics that we seek to incorporate in our proposed DLNSEO metrics:

- Are comprehensive and capture the entire process
- Include the ability to drill down to more detailed information and pinpoint causes
- Are useful in charting progress toward goals
- Do not produce unintended consequences
- Provide early warning in a timely fashion
- Are useful in evaluating the impact of innovations or changes in policies
- Allow comparisons across organizations and time
- Are consistent and reliable
- Do not have costs that outweigh the benefit
- Limit the number of signals to top leadership
- Avoid misleading aggregation

Note that these properties are not mutually exclusive. For example, a metric that permits drilling down to identify causes of problems also serves as an early warning.

Stakeholder interviews

Two of the important characteristics of metrics we noted include limiting the number of signals to top leadership while providing the ability to drill down for more details. Language readiness involves leaders from a variety of commands, each of which has a unique perspective and authority over components of language readiness. Because of

this, we interviewed stakeholders at various levels of leadership and with different oversight authority so that we could understand the types of metrics that would be the most useful to them.

Our interviews included the current and former Deputy Assistant Secretary of Defense for Readiness (DASD (R)—Dr. Laura Junor and Dr. Samuel Kleinman, respectively), and representatives from the Joint Staff, each service's foreign language office, Special Operations Command (SOCOM), the Office of the Under Secretary of Defense for Intelligence (OUSD(I)), and DLNSEO.

Throughout these interviews, a number of common issues were raised. Both the current and former DASD (R) expressed concerns about a "hollow force" and the need to retain the language capabilities the department has worked so hard over the previous decade to build. The current DASD (R) indicated that she was concerned with losing these capabilities as the department downsizes, and she felt that metrics should be established that would help to ensure that the department manages and retains servicemembers with LREC skills. The former DASD (R) echoed these same concerns and also recommended that we pay special attention to the retention of servicemembers with the highest levels of proficiency, which take the longest to train and are therefore the most costly to replace.

Representatives from the services' foreign language offices indicated that one of their most important metrics to track is the number of language requirements filled with servicemembers with the right level of proficiency, referred to as "fill." The Air Force acknowledged that the current fill rate for its language-coded billets is low and thought an important metric would be to see this rate increase.

Representatives from the Joint Staff (J1) echoed the importance of metrics to track the fill rate of language requirements and indicated that DLNSEO's reported language fill rates are alarmingly low. They wondered whether the low rates were the result of an insufficient number of servicemembers with the right skills or a function of the way DLNSEO measures fill. The Director, SOF Language Office, SOCOM, also expressed concerns with the way DLNSEO measures language fill rates. As we discuss later, we conclude that the low fill rates generated by LRI are caused by the method used in its

calculation; we make recommendations for a different—and, we believe, more effective—way to measure it.

While many of the same issues were expressed by representatives from OUSD(I), they offered additional concerns. In particular, they felt that the inclusion of servicemembers who self-profess language proficiency, but who do not test, is misleading in statistics of the proportion of servicemembers with any language proficiency. They understand the importance of identifying these servicemembers so that they could be called on in time of emerging requirements since they could either be immediately deployed if they test at the right level of proficiency or be enrolled in training to enhance their proficiency. They urged that they not be included, however, in metrics regarding proficient servicemembers since their true proficiency is unknown.

Consistent with the findings of our literature search, these representatives also urged against aggregation of metrics of language proficiency. We concur with their recommendation, as we discussed previously.

We turn now to our proposed metrics.

This page intentionally left blank.

Current language readiness metrics: fill and FIT

Our first proposed metrics are language readiness fill and FIT metrics. Both are widely used within and across DOD as measures of readiness. As noted earlier, fill is the number of people filling billet requirements in a particular unit, however that may be defined (e.g., battalion, ship, mission, or OPLAN). So, if 95 people are assigned to a unit with a requirement for 100, the fill rate would be 95 percent.

FIT gives a measure of how well the people assigned to the unit satisfy the requirements, in terms of paygrade, skills, or any characteristics considered to be important to the mission. In the above example, while almost all of the billets have someone assigned to them, a properly constructed FIT metric would measure whether many of those assigned are too junior or do not have the right training. Combined, then, fill and FIT provide quite a bit of information about the number and qualifications of people filling requirements and, therefore, are useful in combination as measures of personnel readiness.

Fill and FIT—in combination—are the metrics we propose that DLNSEO should use to measure the ability of the *current* inventory to satisfy steady-state and contingency requirements that depend on language capabilities. We suggest replacing the method currently used in LRI with the one we propose here. As we will show, our method provides a more accurate assessment of current language readiness, and it addresses the concerns raised in our stakeholder interviews about the way this metric is currently calculated in LRI. If our measures are adopted, we recommend that LRI continue to provide flexibility that allows each combatant command (COCOM), service chief, and so on, whom we refer to henceforth as the "user," to set his or her own priorities and strategies for achieving readiness for the requirements under his command.

Before we describe our proposed measures of fill and FIT, we need to briefly describe how the LRI currently measures language readiness,

which is best defined as a combined measure of fill and FIT, albeit a conservative one.[9] According to the LRI manual [26], "The Services, DLO, and Joint Staff have the ability to compare and match linguist assets to specific COCOM requirements." Their criterion for a match, which they refer to as "Requirements Asset Matching," is defined as a 100-percent match in all of the following: (a) Language Type, (b) Service Requested, (c) Gender, (d) Language Skill (writing and regional expertise is not turned on), (e) Grade (includes one grade up and one grade down), and (f) Security Clearance.

LRI users may ignore various attribute criteria in order to achieve a higher level of FIT. For instance, a COCOM may want to disregard the service requested, assuming that, if there aren't enough matched personnel in one service, matched servicemembers in other services can be substituted.

Redefining fill and FIT

One of our most important points of departure from the current LRI method is that LRI only allows for a 0/1 indication of a match between a person and a requirement; if a perfect match is not made on all specified attributes, the person is regarded as not matching the requirement. In contrast, we specify that requirements have a target level of qualification, but people with qualifications above and below that level—though at least at or above some specified lower threshold—contribute some to the fulfillment of that requirement. For instance, in our revised methodology, a requirement may indicate the need for servicemembers to have an ILR level 2 reading proficiency in a language (target level), but a user may specify that members with an ILR level 1 or 1+ are less desirable but still acceptable, and those scoring 0 or 0+ in reading do not contribute anything to the requirement (in this case, the minimum threshold would be level 1).

As we noted previously, fill is measured at a specified aggregate level, which, for simplicity, we refer to henceforth as a "unit." For our purposes, we refine it to mean the percentage of language requirements in a unit that are filled by servicemembers who meet some minimum

9. For simplicity, we refer to the LRI measure as a measure of FIT.

threshold of proficiency (and other attributes, as we describe later) in the required language. For example, if there is a steady-state requirement for five Farsi linguists with ILR level 2 reading proficiency in a unit, and four are on board who meet the minimum threshold, which we will define here as having at least a level 1 reading proficiency, fill would be calculated as 4/5, or 80 percent.

Our definition of language FIT is that it is a measure of the degree to which individuals filling language requirements match the various components of the requirements, while accounting for imperfect matches. Referring to the example above, while there are four people who satisfy the minimum threshold of reading proficiency, some or all of them may not be at the required level 2 proficiency, which would be reflected in FIT.

Criteria to determine personnel FIT typically include characteristics that have substantial impact on the readiness of that unit, such as occupation and paygrade, and especially for our purposes, language proficiency. FIT is measured first at the individual level and may be aggregated across those individuals associated with a particular unit. When measured at the aggregate level, it is the average FIT of the individuals belonging to that unit. By definition, FIT can never exceed 100 percent because no one can possess more than 100 percent of the target qualifications.

Requirements can be one-dimensional (i.e., based on just one characteristic, such as language proficiency) or multidimensional (i.e., based on language, paygrade, gender, etc.). We discuss each in turn.

One-dimensional requirements

When matching individual capabilities to some aspect of a language requirement (e.g., reading proficiency), a user assigns a score between zero and one to represent the degree to which a servicemember matches that attribute of the requirement. For our metric, the value assigned for a match can be any number between a perfect match (score = 1) and no match (score = 0). Using our Farsi example, the user might consider that individuals with a reading proficiency level of

- 1 or 1+ are a somewhat acceptable alternative to someone with a level 2 proficiency, and assign these individuals a score of 0.5;
- 0 or 0+ do not satisfy the minimum threshold and would therefore be awarded a score of 0; and
- 2 or higher would be assigned a score of 1.

As we noted, this is one important way in which our proposed FIT metric differs from the measure of requirements asset match that is currently calculated in LRI; the latter assigns only values of 0 (not an exact match) or 1 (an exact match).

Continuing our Farsi example, *everyone* in the unit would be ranked on their Farsi reading score (1, 0.5, or 0), and the sum of the top five scores would provide the numerator of our FIT metric. We illustrate this calculation in table 1, with a hypothetical example in which two of the top scorers have a score of 1, and three have scores of 0.5 each. Because this is a one-dimensional example, a person's score is also his or her FIT.

Table 1. Example to illustrate measuring FIT for single attribute

	Proficiency level	Score
Person 1	2	1.0
Person 2	2	1.0
Person 3	1+	0.5
Person 4	1+	0.5
Person 5	1	0.5
Sum		3.5
FIT		3.5/5 = 70%

The numerator of the unit's FIT metric is the sum of individuals' FIT scores, or 3.5, and the denominator is the number of requirements, 5 in our case, resulting in an overall FIT score of 70 percent and a fill equal to 100 percent (since all five incumbents have at least the minimum measured proficiency of level 1). In contrast, LRI would indicate that only the first two individuals satisfy the requirement, if the user ignored all other characteristics (e.g., paygrade, security clearance), resulting in a calculated match of 2/5, or 40 percent.

In the table, we incorporate a tool used to represent fill and FIT—a stoplight dashboard. The dashboard groups ranges of fill and FIT, and associates the measures with categories of readiness. For instance, we specify that FIT metrics of 80 percent or higher are considered to fall into the "ready" (or "go") category and so are shown in green; metrics of 50 to 79 percent are categorized as "marginal," represented by "caution" yellow; and 0 to 49 percent metrics are categorized as "not-ready" and are represented by a red stoplight. Note that these ranges are arbitrary, and we use them for illustrative purposes throughout this section only.

The FIT measure is sensitive to the scores given to partial matches. Suppose that a user specifies that a proficiency score of 1+ has equal value to a score of 2. In the above example, the sum of the FIT scores would become 4, and the overall (aggregate) FIT would then be 4/5, or 80 percent, which would indicate that FIT is in the green zone, but fill would remain unchanged.

Multidimensional requirements

Requirements are often multidimensional, with most language requirements indicating proficiency in two or more modalities. For these requirements, the same scoring scheme would be used for each modality as described above, assigning two scores to each person in the unit.

There are several ways the partial scores can be combined across the two proficiency attributes of the requirement to produce a measure of FIT.[10] We select the harmonic mean (HM) for this purpose, which we describe next.

Harmonic mean

A common practice in calculating an aggregate measure is to use a simple average, or arithmetic mean. However, simple averaging implies that a high score on one attribute can substitute for a low score on another. Earlier we cited the example provided in [24] that describes the problems that can arise with this type of aggregation,

10. For example, one could use the arithmetic or geometric mean.

and we noted that these authors recommend using the HM to correct for such cases. Unlike the arithmetic mean, the HM is more heavily weighted toward low values in a set of numbers; as a consequence, it does not assume perfect substitutability of all inputs.

The HM is defined as the reciprocal of the arithmetic mean of the reciprocals of a set of data. Weights can be assigned to various attributes to indicate that some characteristics are more important than others, resulting in a weighted HM. The formula for the weighted HM is shown in equation 1:

$$HM = \Sigma(w_i)/\Sigma(w_i/c_i) \quad (1)$$

where w_i and c_i are attribute weights and scores, respectively, for full/partial matches of an individual's attributes (i), such as listening or speaking proficiency. Note that the weighted HM is simply the HM if all weights are equal to 1.

Because of the way HM is calculated, in any set of numbers, if at least one approaches 0, regardless of all the others, HM will also approach 0.[11] This makes the HM desirable when aggregating the average of items in which one or more components are so valuable that their absence renders the remaining components useless (e.g., no one has any foreign language proficiency in a unit's language-coded billets).

The HM is best illustrated by an example in which we apply slightly different scores than used in the one-dimensional example. In this case, the requirement specifies a level 2 proficiency in both listening and speaking (i.e., proficiency at the 2/2 level). As before, scores are assigned to each attribute to represent the necessity of that modality.

For our example, we specify that a proficiency score below 1+ is unacceptable, and is given a value of 0. An example of why a user might specify such a low score to either of these modalities is one in which it is determined that either the mission would likely not succeed if

11. The HM is undefined for any set of numbers for which one or more is equal to 0. In the limit, however, as any number approaches 0, the HM approaches 0, which is what we use for cases in which one or more values are equal to 0.

servicemembers were not able to communicate with locals at least at the 1+ level in both listening and speaking or the individuals would be in significant danger if they were less proficient. Proficiencies of 1+ are given a score of 0.5, and all scores 2 and above receive a score of 1.0.

Further, we assume that the requirements for listening and speaking proficiency are equally important, so both have a weight of 1. We calculate both the arithmetic mean and HM for five hypothetical people based on their speaking and listening proficiency scores in a unit that has a requirement for five servicemembers with this level of proficiency. Table 2 shows the results, using stoplight colors to represent the degree of match of individuals' characteristics to requirements.

Table 2. Example of how to score multiple attributes[a]

	Attribute		Score		FIT		
	Listening	Speaking	Listening	Speaking	HM	Arithmetic	LRI
Person 1	2	1	1.0	0	0	50%	0
Person 2	2	1+	1.0	0.5	67%	75%	0
Person 3	2	2	1.0	1.0	100%	100%	1
Person 4	1+	1	0.5	0	0	25%	0
Person 5	1+	1+	0.5	0.5	50%	50%	0
Overall FIT					43%	60%	20%
Fill					60%	100%	20%

a. For our example, we assume that listening and speaking proficiency are of equal importance and are equally weighted (i.e., w = 1).

There are several items to note in this example. First, the first person has a speaking proficiency of 1, which we specified was unacceptable; therefore, he/she received a score of 0 for that attribute. Because the score on that attribute is zero, the resulting HM is zero, and hence that individual's FIT is in the red zone.

In contrast, note that the arithmetic mean for the first person is 0.5 (yellow zone). This person does not satisfy the requirements

according to the HM, while the arithmetic mean would indicate that he/she does, but to a limited extent.

Similar to the first person, the fifth person has an arithmetic mean of 0.50, which means that the two are equivalent in terms of FIT based on the arithmetic mean. Unlike the first person, however, the fifth has an HM of 0.5, because the fifth person (unlike the first) satisfies the minimum proficiency requirement in both modalities. This example illustrates how the HM would discriminate between the acceptable and unacceptable, while the arithmetic mean does not.

On the aggregate level, three people satisfy the minimum criteria, for a fill rate of 60 percent. The HM, our metric of overall unit FIT, is 0.43, or 43 percent, indicating that the average level of FIT in that unit is in the red zone. If instead, the arithmetic mean were used, the fill rate would be 100 percent and the FIT would be 60 percent. Finally, using LRI's methodology, only one person matches on the two modalities; therefore, the calculated requirements asset match is 20 percent. We also indicate this as the fill rate because LRI does not differentiate between fill and FIT.

Our example shows that, of the three ways described to calculate readiness, the arithmetic mean is likely an overestimate, LRI's metric is likely an underestimate, and our metrics are in between these two extremes. Because the arithmetic mean is too generous, we drop it in our ensuing development of our metrics and how they relate to LRI.

Incorporating other attributes

In the previous examples, we measured FIT based only on listening and speaking proficiency attributes. It is likely that additional attributes would be part of most requirements, and this is a consideration in the metric used in LRI.

Some language requirements might include a third modality or a regional expertise, and most would specify non-language characteristics, such as paygrade, LREC community (LP, Special Operations Forces, General Purpose Forces),[12] and security clearance. In our

12. Throughout the remainder of the paper, we use the term *LREC community* to refer to these broad functional groups.

approach, each attribute would receive a score from a user that reflects the degree of substitutability and importance placed on it.[13]

Using a multiple-attribute example, which more closely approximates how LREC requirements will be stated according to [9], we demonstrate the superiority of our FIT to the LRI method. Again, we use a hypothetical example in which there is a requirement for five E5s with 2/2 proficiency in listening and speaking who possess a secret security clearance. We assign E4s a score of 0.7, E5–E6s a score of 1.0, and E7s a score of 0.5; proficiency scores below 1+ receive a 0, scores of 1+ receive a 0.5, and scores of 2 and above receive a 1.0. Finally, no security clearance is awarded a score of 0.3, a reflection of the fact that obtaining a security clearance may take some time, but could be expedited in certain circumstances, while any level of security clearance is given a score of 1.[14] We illustrate the example in table 3, by selecting the top five scorers in that unit, according to the HM.

Table 3. Comparison of FIT with multiple attributes

	Attribute				Score				FIT	
Person	Listening	Speaking	Pay-grade	Clearance	Listening	Speaking	Pay-grade	Clearance	HM	LRI
1	2	1	E5	Yes	1.0	0	1.0	1.0	0	0
2	2	1+	E4	No	1.0	0.5	0.7	0.3	52%	0
3	2	2	E7	No	1.0	1.0	0.5	0.3	55%	0
4	1+	1	E7	No	0.5	0	0.5	0.3	0	0
5	1+	1+	E6	Yes	0.5	0.5	1.0	1.0	67%	0
Overall FIT									35%	0
Fill									60%	0%

13. By importance, we mean both within-attribute and across-attribute; the latter is reflected in weights given to each characteristic. For simplicity, we assign equal weights to all attributes for our examples.

14. Not all servicemembers may be eligible for a security clearance, so an additional attribute could be citizenship. For simplicity, we do not include that in this example.

In our example, using our criteria, three of the five people satisfy the minimum requirement for each attribute, resulting in a fill rate of 60 percent and a FIT of 35 percent. In contrast, LRI would indicate that no one matched the criteria, resulting in a match rate of 0 percent.

Looking at the individuals our metric includes, it is clear that, while no one satisfies all of the attributes at the level specified, each does contribute some measure of readiness to that unit. For instance, the second person has the required listening proficiency, is junior to the specified paygrade, is only a 1+ in speaking, and lacks a security clearance. Obtaining a clearance may not require much time, so that person may be able to be deployed for that mission in a timely manner. Further, his or her deficiencies in paygrade and speaking proficiency are not that substantial, or else we would have specified a lower score for a an E4 and 1+ proficiency. These deficiencies are reflected in a fairly low FIT of 52 percent, but it signifies that this person has at least the minimum specified attributes.

Our example confirms the concerns expressed by some of the stakeholders we interviewed; LRI underestimates language readiness. In contrast, our metrics of fill and FIT, in combination, allow users to understand the extent to which the requirements are not filled at all (fill) and the extent to which they are filled with "the right people" (FIT). In addition, they allow users with different perspectives on the degree of substitutability of attributes to specify the importance they place on different levels of each attribute.[15]

Before discussing how our metrics of fill and FIT can help DLNSEO achieve its vision and mission, we discuss one other significant difference between our approach to measuring language readiness and that used in LRI and by DLNSEO in its current LREC metrics.

Measures of proficiency

In terms of identifying proficient servicemembers, LRI considers people who self-profess but who have never tested to have the same

15. The relative importance of each attribute has a significant impact on the estimation of fill and FIT, and we believe it is an important topic for future research.

proficiency as servicemembers whose last DLPTs were taken the previous week or those whose last DLPTs were taken five years ago. Some of the stakeholders we interviewed expressed concerns about DLNSEO's and LRI's treatment of all three categories of servicemembers' equivalency in terms of proficiency, and we concur; we believe it is a biased overrepresentation of the true capabilities of the total force and could, in fact, have substantial readiness implications.

This conclusion is based on several factors. First, personnel at DLNSEO have told us that, when estimates are compared with actual test results, there is ample evidence that people tend to overestimate their language proficiency. Self-assessed proficiency, therefore, is unreliable, and it risks the success and safety of a mission to assign someone to a billet that requires a minimum of 2/2 proficiency based on self-assessed proficiency only. In addition, if the language is on the Emerging or Immediate investment SLL, there is little justification for servicemembers not testing if they really are proficient because, if they scored at least at the 2/2 level on the DLPT,[16] they would become eligible for FLPB.

Since the services generally require all LPs to retest annually, and since FLPB requires annual recertification, we recommend that language proficiency be differentiated by whether (1) the most recent test is within, say, 395 days, or (2) the last test taken in that modality is more than 395 days before the current date, or (3) it is self-reported only.[17] We illustrate later how this attribute would be included in our measures of fill and FIT.

16. Some servicemembers receive FLPB for 1/1 proficiency if they satisfy their service's exception to the policy, such as being assigned to a language-coded billet or being a member of the Special Operations Forces (SOF) or in support of a SOF unit.

17. To be eligible for FLPB, servicemembers must retest annually. The rules state, however, that the certification period ends on the one-year anniversary of the first day of the first month after the certification date [27]. This would permit a maximum of 395 days between tests. Some organizations use other criteria for determining whether a test is current. For instance, OUSD(I) uses a three-year cutoff for its definition of current tests. As we note in [28], NSA requires civilian employees to be at least at a 3/3 level, and to retest every three years.

Fill and FIT applications

There are two general uses for our fill and FIT metrics to assess readiness and thereby fulfill DLNSEO's requirement to "establish a language readiness-reporting index" [4]. One use is in measuring the actual fill and FIT of already established units, which is what we illustrated with our earlier examples. We need to elaborate on their use for these requirements. First, for current requirements, the fill and FIT should be calculated based on servicemembers serving in that unit. LRI does not include information about the unit of each member, but DMDC's Active Duty Manpower File (ADMF) data do, and we recommend that LRI include this field. That information is imprecise because servicemembers may be on Temporary Duty Assignment (TDY), but the user can account for this in assessing the acceptability of the calculated fill and FIT rates. For instance, if a COCOM wants units to have a fill rate of at least 90 percent, he may conclude that units for which the current fill rate is below 93 percent do not meet that threshold when accounting for servicemembers on TDY.

The other issue is related to the fact that, because it is not possible to identify servicemembers filling particular billets, our method for calculating the fill and FIT of current units includes all servicemembers in a unit; anyone with an HM greater than 0 *could* be filling a language requirement, but may in fact not be, or those with an HM of 0 may be filling a requirement, but we have no way of knowing which is the case. This means that the calculated fill rate could be greater than 100 percent, which does not necessarily mean that more servicemembers are filling the language requirements than are necessary. Our assessment is of language requirements only. These additional servicemembers with relevant language proficiency may be filling other, non-language requirements, and their addition to the unit may or may not enhance the language readiness of the unit. An example would be a unit that requires four General Purpose Forces (GPF) enlisted servicemembers with a 2/2 proficiency in Spanish. Because there are so many GPF who have tested proficiency in Spanish, it is possible that there are far more than four enlisted GPF in that unit who satisfy that requirement, but most are not using their language skills as part of their duties.

The other use for our proposed metrics is in assessing the potential fill and FIT of a contingency operation. In general, the principles are the same for both, but contingency fill and FIT require some modifications. Next, we discuss this methodology for single and multiple contingencies in turn.

Single-contingency operations

Assume that a particular OPLAN's requirements are specified in terms of proficiency in listening and speaking, paygrade, and gender. For this example, we assign a score to proficiency based not only on the level of proficiency, but on the date the last DLPT was taken, or if the proficiency is only self-reported. In addition, LRI includes a field for each servicemember that indicates whether the individual is available for assignment. It is our understanding that the criteria for availability vary by service and are not well documented. We recommend that DLNSEO revisit the way the services fill in this variable and attempt to impose some uniformity, if possible. We noted two properties of good metrics: (1) they allow comparisons across organizations and time, and (2) they are consistent and reliable.

For now, however, we simply note that the first step is to remove servicemembers from the available pool that are categorized as unavailable. If this field includes detailed categories for the reason, the user might be able to be more specific about which reasons would eliminate a member. But, if approximately the same proportion of servicemembers are consistently unavailable for a contingency assignment, it matters more that the user has the right number of servicemembers available, and not necessarily the exact servicemembers who would be available in the case of a contingency operation.

The next steps are required to calculate the potential FIT for that contingency (e.g., OPLAN, or CONPLAN). First, a desired minimum fill rate is selected, with the user giving consideration to other competing requirements and the importance of that contingency operation. The user then specifies the scores that he or she would like to assign the various attributes of the requirements for that contingency, and these scores are used to calculate the HM for every member with some documented proficiency in that language. These servicemembers are

then sorted in descending order of HM, and the top scorers are selected until the specified fill rate is achieved. For instance, if the fill rate specified is 90 percent of a requirement for 100 servicemembers, servicemembers with the 90 highest HMs are selected.[18] Their HMs are then averaged, using the arithmetic mean, to derive a potential "best case" FIT for that contingency.

We refer to this as a "best case" FIT because these calculations depend on the scores and weights ascribed to each attribute by the user, so they reflect that organization's acceptable rate of substitutability and relative importance of attributes. Different weights will result in different servicemembers selected and different levels of FIT. In addition, if the contingency should become operative, it is the sole purview of each service to select their servicemembers for requirements, regardless of the "what if" exercises conducted by other commands. The services must select servicemembers for assignment based on information that is often not readily available in LRI, and even some servicemembers selected in a "what if" by the chief of that particular service may no longer be available if the contingency becomes operable.

The FIT calculated from this exercise may be at an unacceptably low level, which would indicate a lack of readiness for that particular contingency. Or it could be that the fill rate was set too high, in which case a lower fill can be specified and a new FIT calculated.

The HM is also calculated for each attribute of the requirement, by service, to determine where there are deficiencies, as we illustrate in a hypothetical example in table 4. We assume that the user specified an overall fill rate of 90 percent. Note that we include a score for the date of the last DLPT tests; tests taken more than 395 days prior are given a lower score than those that are more current.

In this example, with a specified fill rate of 90 percent, the overall FIT is 79 percent. Notice, however, that the fill rate for each service varies, from a low of 84 percent in the Navy to a high of 94 percent in the Army. Thus, the Navy can satisfy only 84 percent of its requirements

18. This could include servicemembers with an HM of 0.

because that is all of the members in that service who have an HM greater than 0, while the Army can satisfy 94 percent of its requirements. In this example, we allow individual services to exceed 90 percent fill in order to achieve the overall 90-percent target.

Table 4. Dashboard results for a contingency operation: average attribute FIT by service

Service	Paygrade	Gender	Listening	Speaking	When tested	FIT	Fill
Army	0.85	0.64	0.78	0.49	0.75	74%	94%
Air Force	0.77	0.92	0.68	0.57	0.79	83%	88%
Marine Corps	0.59	0.68	0.82	0.48	0.64	77%	92%
Navy	0.61	0.93	0.59	0.59	0.62	80%	84%
All	0.75	0.85	0.72	0.52	0.70	79%	90%

Referring to table 4, the lowest attribute FIT scores are in speaking, and the lowest of these are in the Army and Marine Corps. This is an indication that all services, and these two more so than the Air Force and Navy, need to enhance the speaking proficiency of servicemembers in this language in order to have the capabilities to satisfy this particular contingency. And the low average FIT score for testing indicates that relatively few of the servicemembers in any service are testing on an annual basis, especially in the Navy and Marine Corps. Knowing this, these services might want to establish policies that encourage more frequent testing, or determine whether there is a deterrent to servicemembers testing more frequently.

As this example illustrates, these calculations, and the dashboard presentation of the results, allow users to know whether there are sufficient capabilities to satisfy contingency requirements, and, if not, where the deficiencies are in terms of service, paygrade, modality, and so on. This information is useful in determining what type of remediation is necessary to rectify deficiencies, but, as we describe in the next section, understanding more precisely why the deficiencies exist may require drilling down to more specific metrics. Our fill and FIT dashboard only tells the user that there aren't enough people with the required language skills in the specified paygrades; it does nothing to indicate why this is so.

Some remediation efforts will take less time than others. As we noted earlier, because some language proficiency takes a long time to attain, the minimum threshold for the red zone for modality FIT metrics may need to be set much higher—say, at 85 percent. It is beyond the scope of this study to determine what the right thresholds should be, but we recommend this for further study.

Multiple-contingency operations

Suppose a user wants to know whether there are sufficient personnel in inventory to meet the requirements for more than one contingency or OPLAN that require the same language. In that case, for each OPLAN selected, the user would specify (1) its priority, (2) its desired fill rate, and (3) scores for each characteristic.

The same steps as before are conducted, with an HM for each person calculated for each OPLAN. Those with the highest HM for the OPLAN with the highest priority would be selected until the desired fill rate is achieved. Remaining servicemembers with the highest HM for the OPLAN with the next highest priority would then be selected until that desired fill rate is achieved, and so on, until all OPLANs have been matched with the available personnel. The FIT for each OPLAN would then be displayed in a dashboard for the user to see the FIT for each. If the FIT is too low for any one OPLAN, the user can rerun the exercise with different levels of fill and scores.

Table 5 shows a dashboard that might result from such an exercise. Note that we also include additional statistics that help to indicate the distribution of HM scores. For instance, we include the lowest HM of each member selected for that OPLAN, as well as the 25^{th} percentile HM, which means that 25 percent of those selected have an HM of that value or lower. The HM quartiles (25^{th}, 50^{th}, and 75^{th} percentiles) could be included in every dashboard for both current and contingency requirements, in cases where there is a large enough number of requirements; it does not make sense to report quartiles for units or OPLANs that have fewer than 25 or so requirements.

Again, if the FIT is too low, the user may want to rerun the simulation, using different fill rates or applying different scores and weights to

the attributes. For instance, referring to table 5, with the specified scores, weights, and OPLAN fill rates, OPLAN C has a very low FIT rate of just 34 percent, while OPLAN A, with the highest priority, has a fill rate of 85 percent and a FIT of 80 percent. On one hand, if the FIT for OPLAN C is unacceptably low, the user may want to set a lower fill rate for OPLAN A or B to see what effect that would have on the FIT of OPLAN C. On the other hand, if the FIT of 80 is already unacceptably low, this would not be a reasonable option, and action may be necessary to increase the FIT for OPLAN A.

Table 5. Multiple OPLAN dashboard

OPLAN	Priority	Require-ments	Fill	FIT	Lowest HM	25th percentile HM
A	1	125	85%	80%	12%	20%
B	2	110	85%	72%	22%	31%
C	3	40	80%	34%	6%	11%

Additional fill and FIT metric issues

Language readiness requirements may require proficiency scores in all three modalities—reading, listening, and speaking—or, in many cases requiring GPF, just listening and speaking. However, service-members most often test in two of the three modalities, and reading and listening is the most likely pair. In particular, according to the DLPT data provided to us by DMDC, in the 12-month period of June 2009 through May 2010, of the almost 45,000 modality test pairs or triplets in the same language taken that year, almost 96 percent were in two modalities only, and, of those, 99.5 percent were in reading and listening. The remaining 4 percent were tests in all three modalities. As we noted in [28], some of the services base FLPB pay on the two lowest tested modalities; since speaking is typically the lowest scoring modality of the three, servicemembers have a disincentive to test in all three, and speaking specifically.

The absence of a test score in a modality does not necessarily mean that the person has no proficiency in that modality, only that it has

not been measured.[19] If a score of 0 is assigned to everyone who has not tested in that particular modality, but has tested in at least one modality, those individuals would have an HM of 0. In the case of current requirements, they would not be counted toward achieving fill or FIT, and they would not be selected for a contingency.

This approach would result in an underestimate of language readiness, however, since it seems reasonable that those who test at, say, a level 2 or higher in listening in a language most likely have some proficiency in speaking in that language. The conclusion is less robust for those who test only in listening and speaking because it is possible to have no ability to read a language but to be able to converse fluently.

Because so many servicemembers test in just reading and listening, while many requirements will specify speaking or all three modalities, we provide some suggestions in the appendix for assigning a score to missing modalities.

19. The same is true for those who have proficiency in a foreign language but neither test in any modality nor self-profess. Unlike these servicemembers, however, those who identify themselves as having some proficiency in any modality are more readily identifiable.

Early warning and other metrics

As we described previously, one of the primary purposes of the metrics we propose is to determine whether there are enough servicemembers with the right level of language proficiency to meet current and contingency requirements. Our proposed fill and FIT metrics help determine the extent to which the current force can meet current and contingency requirements, but they do not provide insight into why deficits may exist or why fill or FIT is declining in a language.

In addition, because they are based on *current* inventory, our fill and FIT metrics are not useful in determining the ability of the *future* force to meet these requirements. So, in addition to requiring metrics that drill down to more detailed information and pinpoint causes of current deficiencies, other metrics are necessary to determine whether there will be sufficient servicemembers to meet these requirements in the near term to the longer term. These two types of metrics overlap a great deal and, unlike measures of fill and FIT, they do not depend on requirements to calculate, but requirements do help to inform them.

In this section, we describe these additional metrics, which vary not only by their purpose, but in their targeted audience; the right people must be made aware of the problems identified so that remedies can be put into place by those who have responsibility for the identified problem. For instance, DLNSEO has oversight over FLPB, the service chiefs have responsibility for the management of their own personnel, the Army has responsibility for DLIFLC, and so on.

Our proposed metrics do not make up an exhaustive list. Our purpose is to identify the most critical metrics for DLNSEO to track and to provide guidance for properties of additional metrics. We also describe how to present metrics to various audiences, what types of metrics may be useful in discerning whether potential problems exist, and what data are available to calculate the majority of these metrics.

We begin with a discussion of how the fill and FIT metrics described previously can be used to determine whether additional drilling down is necessary to pinpoint the causes of problems. We then present metrics that can be used for early warning purposes, and to monitor the potential for a hollow force, followed by recommendations for additional reports and a recommendation for DLNSEO regarding data management and report generation.

Using fill and FIT metrics

We describe our metrics in terms of a process. The first step in that process is to determine whether shortfalls exist in current requirements, using our proposed fill and FIT metrics. If either metric highlights a deficiency in any attribute (e.g., paygrade, modality, service) of a language, that will help to identify additional metrics that will be necessary to pinpoint specific causes. Because of quarterly phenomena in recruiting and separations, we recommend that, once the Capabilities Based Requirements Identification Process is complete, DLNSEO begin to track these metrics quarterly. DLNSEO can begin to calculate these metrics now, however, for LPs because steady-state requirements have been fully vetted and we believe that the majority of these requirements will be for LPs.

The same scoring should be used for all attributes to ensure consistency, so that trends can be tracked over time for the same language, and so that languages can be compared within the same quarter. Recall that these are two of the properties of good metrics. For instance, to track fill and FIT over time and across services, similar attribute scores should be used for each modality each quarter and for each service.

It is also important that fill and FIT, and in fact most of our metrics, be calculated separately for each LREC community. We recommend this for two reasons. First, DLNSEO is required to monitor certain phenomena related to LPs and FAOs, as we described. The second reason is that aggregating the metrics to the Total Force is misleading because servicemembers in different LREC communities are generally not close substitutes; a GPF cannot fill a requirement for an LP,

and vice versa.[20] While the services may be increasing the number of LPs with proficiency in a particular language on the Immediate list, there may be a perfectly offsetting reduction in the number of GPF with that language proficiency. An aggregated metric that includes all servicemembers would mask this potential problem.

At the very least, then, we propose that separate calculations be made for each component (active, reserve, and Reserve Officer Training Corps (ROTC)/academies), and within each, for each LREC community.

It may be desirable to differentiate the third category further, into language-enabled servicemembers. This would include members of the SOF, or members that individual services refer to as language enabled, and all other GPF.

In addition to tracking by LREC community, we recommended that two categories of fill and FIT be calculated: (1) all current requirements and (2) all current plus one or two contingencies with the greatest requirements or priorities for that language. Clearly, these calculations only make sense if there is a sufficient number of servicemembers in each cell (component/LREC community/language), but for those languages with limited requirements, one measure of fill and FIT may be sufficient.

A substantial quarter-to-quarter decrease in fill and/or FIT in any cell (i.e., language/service/payband) provides an early warning that readiness is declining, in which case DLNSEO would require additional analysis to determine the cause for the decline. That is part of the second step, which we turn to next.

Metrics after calculating fill and FIT

The next step requires the measurement of two types of metrics. One set of metrics should be measured and tracked quarterly, the same as

20. On one hand, LPs are usually not required to speak in a foreign language as part of their official duties, nor are they trained to be interpreters. GPFs that are heritage speakers, on the other hand, while not trained formally as interpreters, may serve in that capacity better than LPs.

fill and FIT, and serve as the indicators of whether substantial changes may be happening in the inventory of servicemembers with specific language skills—the early warning metrics. The other set of metrics will depend on whether there were substantial changes in fill and FIT, or if the early warning metrics indicate a problem. If so, these metrics will drill down to pinpoint the causes of problems. We begin with early warning metrics because they help to determine whether drilling down is necessary.

Early warning metrics

Significant changes in the inventory of servicemembers with certain language skills could occur in the near term to the longer term future because of changes in one or more of the inflows and outflows we noted earlier or because of other issues, such as training backlogs and higher training attrition. These changes may be unique to a particular service, to particular paygrades, or to particular languages. The metrics must help to discern which skills are at risk or are deficient so that the cognizant authority can pursue appropriate remedies.

Unless specific issues arise from the fill and FIT calculations, such as a low FIT in one language for one service only, many of these other metrics are not required to be tracked quarterly. Instead, we suggest that just two types of metrics are necessary for this purpose. The first type of metric is the most efficient way to identify whether the number of proficient servicemembers is declining. Depending on the outcome, additional metrics may be necessary to determine the source of the reduction in proficient servicemembers.

We recommend that DLNSEO measure the inventory of servicemembers who have tested at various levels of proficiency, by language, at the end of each quarter. These levels include the following (and these calculations should be made separately for each LREC community):

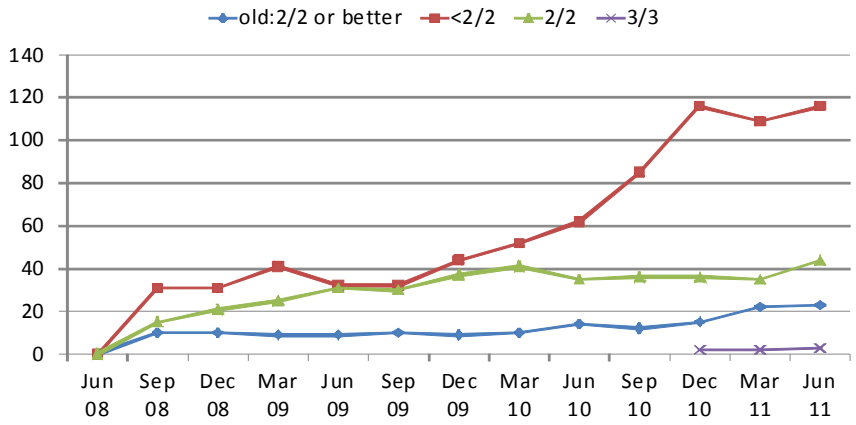
- Current tests,[21] less than 2/2

21. As we noted in our discussion of fill and FIT, reporting the proficiency of servicemembers should differentiate the source and last test date of proficiency. We recommend that the number of servicemembers who have no tested proficiency be excluded from this metric.

- Current tests, 2/2 to less than 3/3
- Current tests, 3/3 or better
- Older tests that are at the 2/2 or higher level.

We illustrate this metric in figure 1 with a hypothetical example. For this language, we tally for each quarter the number of active component (AC) GPF with (a) tests at the 2/2 level or higher, but the most recent tests were taken more than 395 days before the end of the quarter, (b) current tests below the 2/2 level, (c) current tests 2/2 or higher but below 3/3, and (d) current tests 3/3 or higher. For this purpose, we define current as tests that were taken in two modalities within 395 days of the end of that quarter.

Figure 1. Number of AC GPF servicemembers testing in language



Referring to figure 1, the number of servicemembers with current tests at the 2/2 level increased in the last quarter. There was also a substantial and steady increase in the number with current tests below 2/2 between September 2009 and March 2011, but their numbers have remained fairly stable for the last three consecutive quarters. In retrospect, the increase is fairly dramatic, and the reason(s) for it should probably be pursued to help identify the impact of innovations or changes in policies (recall that these are properties of good metrics).

On one hand, if these trends had been graphed each quarter, the increase in the number with any tested proficiency that began in March 2010 would likely have been reflected in increasing levels of fill and FIT if requirements were fairly steady. On the other hand, there are two potential areas of concern that are apparent in this graph: (1) there has been a steady increase in the number who tested at least at the 2/2 level in the past, but have not retested recently, and (2) very few GPF have current tests at the 3/3 or higher level. More in-depth analysis might be required to determine the causes for these potential problems: are they primarily the problem of a particular service or particular paygrade (enlisted versus officers), or can they be traced to other sources?

To pursue the first potential problem, the next step could be to plot, by payband, the number of AC enlisted GPF who have no current test scores but who scored at least 2/2 in that language in the past, as we do in figure 2. This plot helps to determine whether the lack of recent testing is unique to a particular payband (a similar graph for officers would also be required). Note, for instance, that almost half of the GPF without current tests in the quarter ending June 2011 were servicemembers in paygrades E4 through E6.

The next graph helps to drill further to determine whether the problem with mid-grade enlisted GPF servicemembers is common to all services or is an isolated problem (see figure 3). While the numbers are not large in any service, the Army has the greatest number (four GPF) who did not retest last quarter.

Combined then, figures 2 and 3 indicate that, among AC enlisted, it is the mid-career servicemembers who are the most likely to fail to retest annually, and, of these, the Army represents the greatest share. These two findings may be an indication that the failure to retest is caused by deployments, but more analysis may be required to determine whether this is the case, or whether it is because of backlogs in testing or other problems. However, there are relatively few who aren't retesting, so DLNSEO may want to simply wait another quarter to see whether the trend continues.

Figure 2. Number of AC enlisted GPF with old scores 2/2 and above in language

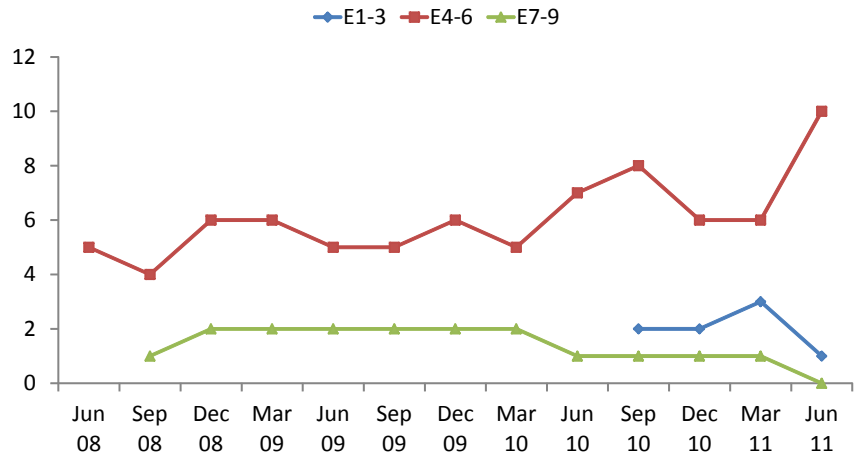
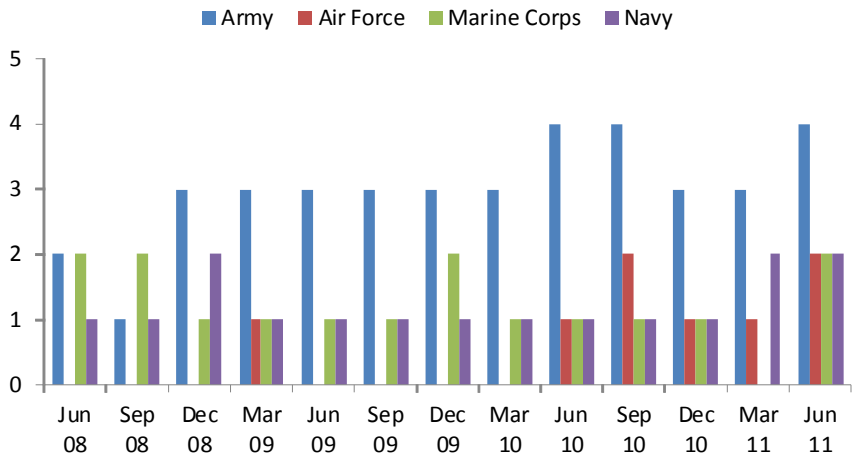


Figure 3. Number of AC E4–E6 GPF with old scores 2/2 and above in language



Finally, referring back to figure 1, the total number of AC non-LPs with any tested proficiency in the language has been increasing fairly steadily over these eight quarters. While this does not provide insight as to the number of gains or losses of these servicemembers, or whether the gains have been proportional across all services and pay-grades, it does indicate that the total number of losses has not

exceeded the gains. The concerns expressed regarding the hollow force, however, require the second set of metrics to determine whether there are potential problems in recruiting, attrition, promotion, or retention.

Hollow force metrics

DLNSEO needs to determine which of the following is true of servicemembers with foreign language proficiency:

1. Are they disproportionately attriting, separating, or being passed over for promotion?
2. Are fewer heritage speakers being recruited?

The quarterly graphs we just discussed provide a good indication of whether the Total Force is losing a substantial number of proficient servicemembers, but they don't answer these concerns directly.

To do this, we recommend that DLNSEO construct the same quarterly graphs that we just described, with separate graphs for each LREC community/payband. For example, one graph would be constructed for all GPF servicemembers in each of the following paybands: (1) E1–E3 (for measures of recruiting and early attrition trends), (2) E4–E6, and (3) E7–E9. The latter two graphs are overall indications of advancement and retention, but they are not specific measurements of each phenomenon.

Once these overall trends are noted, additional metrics may be needed to determine the source of troubling trends. For instance, if the number of servicemembers at the 2/2 level in paygrades E4 through E6 is stable while the number in the more senior paygrades is decreasing, it could be that mid-grade servicemembers are not advancing to senior paygrades at the same time that senior proficient servicemembers retire.

For these more specific metrics, DLNSEO needs to calculate the attrition, separation, and promotion rates of proficient servicemembers *relative* to *similar* (in terms of occupation and paygrade) servicemembers who are not proficient. The emphasis is on relative and similar; we caution against measuring the absolute rates of attrition,

promotion, and retention of proficient servicemembers, or these rates relative to all nonproficient servicemembers. To do otherwise would yield misleading results.

Consider, for example, the case in which GPF servicemembers who are proficient in a particular language are disproportionately in one service, and within that service, are within a particular enlisted occupation. In fact, this is often the case, at least in terms of the distribution of proficient GPF servicemembers in specific languages across services, as we documented in [28]. Assume further that the service is downsizing, and disproportionately so in that occupation, but proficient servicemembers in that occupation at the end of their second term are separating at half the rate of their nonproficient peers (e.g., 30 percent versus 60 percent). Simply noting that the second-term reenlistment rate has gone down for these servicemembers, or comparing their separation to that of all servicemembers in the entire service who are not proficient, regardless of their occupation, would lead to the conclusion that these servicemembers are not being retained in sufficient numbers. It could be argued that the service needs to retain more of these proficient servicemembers, but this would require that service to deny the reenlistment of almost all nonproficient servicemembers in that occupation in favor of proficient servicemembers, regardless of their ability to perform their occupational duties. Such a requirement would have troubling readiness implications for that service.

Recall also that one of the properties of a good metric is that it avoids misleading aggregation. Hence, we recommend that, in order to determine whether problems exist in the attrition, retention, or advancement of proficient servicemembers, the analysis be confined to relative measures of proficient servicemembers versus nonproficient servicemembers in the same service, payband, and at least LREC community.

Languages to track

Until the CBRIP is complete, we recommend that individual graphs be constructed for each language on the Immediate and Emerging SLL list, and two graphs for the remaining languages: (1) all other

languages on the Enduring list combined, and (2) all other languages on the Enduring list except for Spanish and perhaps a few of the other more prevalent languages. The reason for the second graph is that, as we documented in [28], Spanish has consistently represented about 25 percent of all tests taken for the past several years. Because it represents such a large proportion of all tests taken in Enduring languages, large changes in most of the other languages in the list would be indiscernible if Spanish were included.

After the CBRIP is complete, we recommend that graphs continue to be constructed for languages on the Immediate and Emerging list, and for any language not on one of these lists that either is below an acceptable threshold or for which fill and FIT have decreased for, say, two quarters in a row, or by more than some specified amount within a quarter. Again, the threshold for when more in-depth analysis is necessary is beyond the scope of this project, but we submit that these are important areas for future research.

Drilling down

We have already described cases in which it might be necessary for DLNSEO to drill down to pinpoint the causes of problems that the early warning metrics of fill, FIT, and levels of proficiency reveal. It is not practical to describe every possible metric for this purpose here, but we submit that the choices would be obvious in most cases. For instance, if fill or FIT is sufficient for every service except the Air Force, DLNSEO could first calculate the quarterly proficiency metrics for the past few years for servicemembers of the Air Force, by pay-band, to determine whether the deficiency is across all paygrades or specific to just some. If DLNSEO determines that it is unique to one payband (e.g., E4–E6), it can then dig deeper to determine whether the problem is caused by recruiting, training, attrition, advancement, or retention problems.

Metrics tracked annually

Some metrics may not show enough quarter-to-quarter changes to warrant measuring at that level of frequency. Instead, we propose that these metrics be tracked annually; they not only serve the purpose of early warning but also provide information that is useful in estimating

the full range of costs and benefits of LREC readiness, which we discuss later, that is necessary to establish metric goals.

Specifically, we recommend that DLNSEO calculate various outcomes of servicemembers trained at DLIFLC, such as the average time to train LPs in each language, the percentage of graduates who graduate with DLPT scores above 2/2/1+,[22] the attrition rate of various classes, and so on.

We also recommend that DLNSEO report annually on the number of new accessions who (1) take either a DLPT or Oral Proficiency Interview (OPI) within six months of accession, by service and language, and (2) do not take a test but report some proficiency. These two metrics help to determine the extent to which the services are screening new accessions, and they help to identify the available pool of servicemembers with some basic understanding of various languages.

In [28] we recommended that questions regarding the source(s) of proficiency of servicemembers testing for the first time in a language be added to all DLPTs/OPIs, and that a similar question also be added to get servicemembers to signify how they obtained their most recent proficiency if they are retesting (such as military training, in-country experience, or voluntary education). If these questions are added, DLNSEO could use that information to help address such questions as the effect of FLPB on encouraging servicemembers to learn new languages or to maintain or enhance their proficiency. If these questions are added, DLNSEO should summarize the information that servicemembers provide, such as the frequency of test-takers reporting proficiency from each source, the average proficiency by each source, and which source is most associated with enhanced proficiency in subsequent tests.

We also recommend that DLNSEO conduct an annual review of FLPB payments made in each component, such as how much is being paid for each language, in each service, and so on.

22. These are the current standard proficiency goals for LP graduates of DLIFLC.

Other annual metrics may be useful, as warranted, and as data become available. For instance, once all requirements are established, we recommend that DLNSEO track the number of requirements for each language, at each level of proficiency/LREC community/service, by whether they are steady state or surge, and other characteristics as necessary. LREC requirements will likely not change very often, so it is not necessary to track them more frequently, but it would be useful to have a historical annual record of these requirements.

Summary of metrics

We summarize the metrics we recommend, by their frequency, in table 6. In addition, other metrics will be required on an as-needed basis each quarter, as we discussed.

Table 6. Summary of metrics[a]

Metric	Aggregation	Details
	Quarterly	
Fill and FIT	Each language separately: 1. Current requirements 2. Contingencies with highest priorities or requirements	Currently, only for LP steady state. After FY14, all languages with requirements.
Number of people with current tests <2/2, 2/2, 3/3, all older than 395 days	1. Each Immediate and Emerging language separately 2. All others combined 3. All others without Spanish and other prevalent languages	These should be calculated and graphed. Once CBRIP is complete, only Immediate and Emerging languages and those with fill or FIT below threshold.
Number of people with current tests <2/2, 2/2, 3/3, all older than 395 days	Same as above but separate calculations for paybands. If problems are identified, additional calculations by service/paybands, as needed	These are macro hollow force metrics used to determine if additional drilling down is necessary

Table 6. Summary of metrics[a]

Metric	Aggregation	Details
Percentage of language-proficient members who advance/attrite/retain one pay-grade relative to percentage of similar nonproficient servicemembers	For servicemembers in the same service/payband/occupation	Hollow force metrics to drill down to specific causes of decreasing capabilities, if necessary
	Annual	
Number of new accessions who test within 6 months at the <2/2, 2/2 and 3/3 level	By language and service	These are indications of (1) heritage recruiting efforts, (2) service's testing policy, and (3) available pool of servicemembers with any knowledge of particular languages.
Number of new accessions who do not test but report some language proficiency	By language and service	
Average time to train at DLIFLC	By language	These metrics help to determine if there are changes in the length or effectiveness of training. In combination, they also help in determining the cost of training (in terms of days spent in training) by language and level of proficiency.
Percentage of servicemembers who begin DLIFLC training but do not complete	By language	
Percentage of DLIFLC graduates who test at least at the 2/2/1+ level	By language	
Percentage of DLPT/OPI test-takers who report various sources of proficiency	By language	This information helps to determine the effectiveness of FLPB and training in enhancing and maintaining proficiency.
FLPB payments	By language/service	Helps to track the costs of FLPB and to project FLPB future budgets
Requirements	By language/steady state/contingencies	Requirements will not change frequently, but it is important to keep track of them for historical records.

a. Each metric should be calculated separately for each component, and within each, for each LREC community.

Data

Throughout this section, we have recommended that DLNSEO calculate various metrics that we recognize are currently beyond their ability to do. The primary reason they cannot is that they do not have the data in-house. Instead, DLNSEO relies on outside entities to provide inputs for the various reports they are required to generate. This often results in a lack of consistency in metrics reported across organizations, such as those they noted in the 2010 report required by DODI 5160.70 [3]. DLNSEO has also shared with us its ongoing efforts to reconcile proficiency data in LRI between data provided by DMDC and that provided by the individual services; DLNSEO has had persistent difficulty in understanding which source is right and why they don't agree.

For the purpose of this and previous studies for DLNSEO, we have obtained personnel, DLPT, OPI, and pay data from DMDC, merged the various data sets by servicemembers' Social Security Numbers (SSNs), and calculated many of the metrics we propose. In so doing, we have become very familiar with problems with some of these data, inconsistencies across services, and so on. Even so, we have been able to resolve these problems, and, as a consequence, we have developed a very good understanding of the available data and how to manage them.

Our experiences lead us to the conclusion that, in order to have the most comprehensive strategic oversight of the Defense Language Program and to provide the highest quality reports required of them, DLNSEO needs to acquire the necessary data and calculate these metrics in-house. At the very least, this requires DLNSEO to obtain from DMDC quarterly extracts of the Active Duty Military Personnel Master File and the Reserve Component Common Personnel Data (RCCPDS), to include SSN, service and component, paygrade, primary/secondary/duty occupation, and unit assigned. Other variables, such as gender, race/ethnicity, education, length of service, and time to End of Active Obligated Service (EAOS) for enlisted servicemembers, may also be desirable. To these data, DLNSEO would need to add monthly DLPT and OPI data, monthly FLPB pay data, and perhaps some additional data from the services as needed.

DLNSEO would also need to obtain data on steady-state requirements, until they are added to LRI.

We understand that some of the discrepancies between DMDC and service data result from the fact that the services alter DLPT data for servicemembers based on individual policies. For instance, some services will not allow servicemembers to test in an upper range DLPT until they have scored at least a 3 in a lower range. Servicemembers who do not follow this policy may have their upper range DLPT scores erased until they comply with that policy. However, the DLPT scores from DMDC are consistent in that they are not subject to individual services' policies, which differ across services and even within a service across time. For this reason, we submit that the DMDC DLPT data satisfy the criterion of consistent metrics; therefore, DMDC DLPT data should be the only DLPT data used by DLNSEO to produce metrics and reports.

Managing these datasets, conducting the analysis, and producing the numerous reports should be the responsibility of a dedicated person (or persons) in DLNSEO—a type of "skill manager" that is similar to the service's occupation managers, who manage the entire career path of individuals in occupation groupings within their service. For DLNSEO, the skill manager would be concerned not with occupations but with LREC skills of all servicemembers. Optimal oversight of the Defense Language Program requires the type of intimate knowledge that comes from a robust understanding of these data, including the analysis necessary to track trends and generate reports.

Other uses for these metrics

The metrics we have proposed serve a number of additional purposes. For instance, tracking changes in fill and FIT by language would help inform the SLL. If languages that have a fill and/or FIT rate below some preset threshold are put on a higher priority category on the SLL, the level of FLPB offered for that language would increase, which would provide a signal to the services that there is an insufficient number of proficient servicemembers in that language.

These metrics also serve DLNSEO's requirements for reporting, as we described earlier. If they are calculated quarterly, they will be readily available for these reporting purposes.

We propose one additional report for DLNSEO to generate quarterly that would provide useful signals to top leadership while providing early warning in a timely fashion. Specifically, as we noted previously, DLNSEO has limited oversight of many of the inputs into LREC readiness. We submit that for those inputs outside its domain, it should assemble and disseminate a quarterly report of metrics to leadership, as part of DLNSEO's strategic oversight, so that DOD leadership has timely warning of factors that are under its individual purview.

In particular, we recommend that DLNSEO compile a quarterly report that summarizes fill and FIT calculations, and their assessment of the number of proficient servicemembers in each language. Additional metrics should be included that help to shed light on any concerns raised in previous quarterly reports, or that drill down to provide more information regarding emerging problems. This report should be disseminated to high ranking Defense Language Personnel, such as members of the Defense Language Steering Committee (DLSC).

In table 7, we provide a hypothetical example of how the fill and FIT metrics could be displayed in this report. The report could consist of a number of tables like this one, each indicating the level of readiness for each language, and for each service for that particular language. Separate tables should be created for each LREC community. The languages could be ordered in terms of priority, beginning with Immediate languages first. The actual table would have all languages included, but we include just one for illustration.

Table 7. Table template for reporting fill and FIT quarterly

	Fill (change since last quarter)					FIT (change since last quarter)				
Language	Army	Air Force	Marine Corps	Navy	All	Army	Air Force	Marine Corps	Navy	All
Immediate										
A	83% (+0.4)	83% (-0.5%)	86% (+0.6%)	78% (+0.3%)	82% (+0.9%)	81% (+0.1%)	81% (-0.2%)	74% (-0.5)	76% (+0.8%)	80% (-0.4%)

In addition to reporting the absolute fill and FIT, it would indicate the change, in percentage points, from the previous quarter. While we propose that LRI users should be able to calculate fill and FIT at any time, this table would provide a historical tracking of metrics. DLNSEO might also want to produce an end-of-fiscal-year report in which the year-to-year changes are noted.

Similar tables could be constructed for current tested proficiency, as we show in table 8. We specify reporting current tests at the 2/2 level or higher, but other tables could be included that measure other levels of proficiency.

Table 8. Table template for reporting quarterly

	Number in inventory with current DLPT 2/2 or above (change since last quarter)					Number in inventory with older DLPT 2/2 or above (change since last quarter)				
Lan-guage	Army	Air Force	Marine Corps	Navy	All	Army	Air Force	Marine Corps	Navy	All
Imme-diate										
A	140 (+1%)	200 (-2%)	30 (N/C)	100 (+3%)	470 (N/C)	30 (-3%)	50 (-2%)	10 (N/C)	40 (N/C)	130 (-2%)

In addition to these tables, other tables and/or graphs could be included that provide insight into problems identified in previous quarters or that arose in that quarter, such as decreasing retention. DLNSEO may also want to periodically include a "special topics" table, which would highlight any analysis it has done of issues raised (e.g., in Defense Language Advisory Panel (DLAP) meetings) or of the impact of innovations or changes in policies, in accordance with some of the properties of good metrics.

Language proficiency goals

We have deliberately excluded the establishment of specific goals for our metrics because we believe that setting goals is inadvisable at this time. First, absent the requirements that are still being vetted, there is no way of knowing whether there are enough, too few, or even too

many people with proficiency in specific languages. The one exception could be currently established steady-state requirements, most of which would involve requirements for LPs and perhaps language-enabled servicemembers, but we submit that each service has a vested interest in setting its own goals for these servicemembers, and these communities are closely monitored by their respective service managers; DLNSEO goals for these servicemembers would be superfluous, and beyond their responsibility to achieve.

One could argue that, even absent requirements, it is important to increase the number of GPF who are proficient, or to increase the proficiency of these servicemembers. The problem, however, is that attaining these goals will be very costly—financially and perhaps in terms of readiness. For instance, GPF LREC requirements will likely be mostly for surge and contingencies, and relatively few for steady-state requirements. Futhermore, these servicemembers have primary occupations that, in some cases, require lengthy training. Additional language training, especially for most of the Immediate or Emerging languages on the SLL, also requires formal training that can take a year or more of full-time study. Time spent in language training is time not spent on enhancing their primary occupation skills, and it is not time spent on full duty. To ensure that the GPF maintain the same full-duty experience, and hence the same level of non-LREC readiness, servicemembers would need to have longer service commitments either in the form of longer initial service obligations or higher retention, both of which are costly because of additional incentives required. Further, the services would need to increase their endstrength proportionally in order to fill the steady-state requirements left vacant by those who are in language training. Added to these costs are the financial costs of instructors and infrastructure for the LREC training itself.

The other option is for GPF members to gain proficiency in their own time. FLPB was intended to incentivize such behavior, but a recent CNA study concluded that FLPB has not been successful in this regard [28].

Once requirements are fully established, we submit that research that would validate the establishment of goals is lacking, and the analysis

required is complicated and often takes months. Further, in some cases, the data necessary to conduct the analysis do not exist or are not readily available (e.g., see [28]). For instance, little is known of the consequences of having too few speakers at the ILR level 2 in a language on the ability of a particular unit to perform its duties. Goals need to be established based on criteria, such as an acceptable level of risk that leadership is willing to assume if requirements are not fully met, or fall short by, say, 5 or 20 percent.

Finally, there is no doubt that incremental improvement toward stated goals would be costly. If no requirement exists to justify them, no authority to enforce them, or additional funds to achieve them, it is unlikely that the services will make significant efforts toward achieving these goals.

Hence, we recommend that DLNSEO wait to establish metric goals until the requirements are fully vetted, in FY14. Until then, DLNSEO needs to work to ensure that the data necessary to analyze the costs and benefits of various goals are collected and analyzed so that the full readiness and financial costs of various goals are fully understood when the requirements are available. This recommendation is consistent with two of the properties of a good metric: (1) the costs of measuring the metric do not outweigh its benefit, and we believe the costs of setting metric goals are far outweighed by any current benefit, and (2) they do not produce unintended consequences, which we believe could happen if goals are set without a better understanding of their readiness and financial costs and benefits.

Absent requirements, we submit that the tracking of the metrics we propose will provide ample evidence of the direction of LREC readiness in the Total Force, in specific services, and in languages of strategic importance. The reports that we recommend DLNSEO provide to leadership serve as the strategic oversight required of DLNSEO, and their presentation of successes and concerns can provide valuable guidance to leadership that will help to direct funding, infrastructure, and so on, that will ultimately be required to enhance proficiency when and if necessary.

This page intentionally left blank.

Summary and recommendations

In summary, we propose that DLNSEO develop key metrics that can be calculated when the Capabilities Based Requirements Identification Process is complete as well as metrics that can be calculated now based on the current inventory of personnel with language capabilities. The key metrics to be calculated once requirements are in place include a measure of the current number of servicemembers with any level of proficiency that could fill current and contingency requirements (fill) and a measure of the extent to which these members satisfy the full range of these requirements, in terms of proficiency in all language modalities, paygrade, service, and so on (FIT).

We then propose additional metrics that provide the ability to drill down to more detailed information and pinpoint causes of problems that are identified in fill and FIT calculations, such as in recruiting, training backlogs, and attrition, and also provide early warning that deficiencies in LREC capabilities might arise in the future. Many of these metrics can be calculated now, however, before the conclusion of the CBRIP.

The first set of these metrics that we propose DLNSEO measure is the inventory of servicemembers who have tested at various levels of proficiency in each language at the end of each quarter. Again, these calculations should be made separately for language professionals, FAOs, language-enabled servicemembers, and the GPF. Although declining fill and FIT levels will help determine which languages to focus on in the future, critical languages on the SLL can be tracked now, regardless of requirements.

The second set of key metrics that we propose DLNSEO track quarterly will help determine whether servicemembers with foreign language proficiency are disproportionately attriting, separating, or being passed over for promotion, or whether fewer heritage speakers are being recruited. We also discuss additional metrics that we

recommend DLNSEO track on an annual basis, such as DLIFLC time to train and attrition, as well as FLPB payments.

For all metrics, we caution against those that result in misleading aggregation. As we've noted, reporting the number of servicemembers with any proficiency in any language, including proficiency that is only self-professed but never formally tested, and in languages that are not of strategic importance provides very little useful information.

We also caution DLNSEO against aggregating metrics to determine whether actions are leading to an LREC hollow force. Measuring the retention of proficient servicemembers in isolation, without comparing their retention to their peers, is again misleading and ignores the reality that many of the services are in the process of downsizing. The better metric is whether more proficient servicemembers are leaving relative to their otherwise similar peers, and of special concern is the relative loss of those with proficiency in languages of the greatest strategic importance.

In addition, our metrics purposefully avoid establishing goals. Until the current language requirements process is complete, it is not possible to know whether there are too few, too many, or the right number of servicemembers with language proficiency. Further, goals need to be based on an understanding of the costs and benefits—financial and readiness—of achieving them. We submit that data necessary to estimate these trade-offs are not currently readily available and, in many cases, will be costly to obtain.

We also emphasize the importance of DLNSEO obtaining the necessary data and calculating these key metrics itself. Relying on outside entities to provide inputs for the various reports it is required to generate often results in a lack of consistency in metrics reported across organizations. A dedicated "LREC skill manager," similar to the services' occupation managers, should be appointed to perform these duties, which would include obtaining all of the relevant data from DMDC, the services, DLIFLC, and so on, and to produce the required reports. We believe that only by becoming familiar with these data will DLNSEO be able to provide the full range of support and oversight required of it for the Defense Language Program.

Appendix: Assigning scores to missing modalities

One option for assigning scores for missing modalities is to derive estimates using proficiency scores of members who tested in all three modalities to determine the relationship in the scores between modalities. For instance, statistical regression analysis could be used to estimate (predict) a given modality score from the other two modalities, by digraph. An adjustment may be necessary if the reason that servicemembers who do not test in speaking is that their proficiency in speaking is lower than for those who choose to test their speaking proficiency. If derived estimates (imputed scores) were to be used, FIT reports generated from these estimates should indicate that estimates were used.

Another option is to give partial credit for unmeasured modalities. Rather than estimate a missing score, a user could assign it a score lower than the lowest minimally acceptable score. For example, suppose the (ideal) requirement is for a speaking proficiency score of 3, but a score of 1+ would be acceptable. A user would specify credits of 1.0 and 0.7, respectively. A missing speaking score might be given a weight of 0. This would avoid assigning a score of 0 to the person with the missing score, if someone with an HM as low as 0.5 is acceptable.

This page intentionally left blank.

References

[1] *Defense Language and National Security Education Office (DLNSEO) website,* last accessed Apr. 2, 2012, at http://prhome.defense.gov/RFM/READINESS/DLNSEO/MISSION.ASPX.

[2] U.S. Department of Defense. *Defense Language Transformation Roadmap*. Jan. 2005, last accessed Apr. 27, 2012, at http://www.defense.gov/news/Mar2005/d20050330roadmap.pdf.

[3] U.S. Department of Defense Instruction (DODI) 5160.70. *Management of DoD Language and Regional Proficiency Capabilities*. Jun. 12, 2007.

[4] U.S. Department of Defense Directive (DODD) 5160.41E. *Defense Language Program (DLP)*. Oct. 21, 2005.

[5] U.S. Department of Defense Instruction (DODI) 1315.17. *Military Department Foreign Area Officer (FAO) Programs*. Apr. 28, 2005.

[6] U.S. Department of Defense. *Strategic Plan for Language Skills, Regional Expertise and Cultural Capabilities (2011–2016)*. Feb. 2011, last accessed May 5, 2012, at http://prhome. defense.gov/RFM/READINESS/DLNSEO/files/STRAT%20PLAN.pdf.

[7] Laura J. Junor. "The Defense Readiness Reporting System: A New Tool for Force Management." *Joint Force Quarterly,* issue 39, 4th quarter 2005.

[8] *Defense Readiness Reporting System website,* last accessed Apr. 23, 2012, at http://userguide.drrs.org/LRI/Overview/P02.html.

[9] Chairman of the Joint Chiefs of Staff Instruction (CJCSI) 3126.01 23. *Language and Regional Expertise Planning*. Jan. 2006.

[10] U.S. Department of Defense Instruction (DoDI) 5160.70. *Annual Foreign Language Report*. DD-P&R(A&Q)2272. Jul. 2011.

[11] Solomon W. Polachek et al. "Educational Production Functions." *Journal of Educational Statistics*, vol. 3, Autumn 1978.

[12] Richard E. Just et al. "Estimation of Multicrop Production Functions." *American Journal of Agricultural Economics*, vol. 65, no. 4, Nov. 1983.

[13] Michael A. Salinger. "The Meaning of 'Upstream' and 'Downstream' and the Implications for Modeling Vertical Mergers." *The Journal of Industrial Economics*, vol. 37, no. 4, Jun. 1989.

[14] F. M. Scherer and David Ross. *Industrial Market Structure and Economic Performance*. 3rd ed. Boston: Houghton Mifflin, 1990.

[15] Ali Hortacsu and Chad Syverson. *Why Do Firms Own Production Chains?* Apr. 2009, last accessed Dec. 10, 2011, at http://www2.vwl.uni-mannheim.de/fileadmin/user_upload/thadden/fsem/syverson.pdf.

[16] M. Deru and P. Torcellini. *Performance Metrics Research Project - Final Report*. Technical Report NREL/TP-550-38700. Oct. 2005.

[17] Serhiy Kharytonov. *How To Use Metrics To Improve Project Management*. Feb. 1, 2010, last accessed Apr. 27, 2012, at http://www.eweek.com/c/a/IT-Management/How-to-Use-Metrics-to-Improve-Project-Management/1.

[18] Fern Halper. *A Guide to Building a Metrics-Driven Organization*. A Hurwitz White Paper. 2010, last accessed Dec. 22, 2011, at http://www.hurwitz.com.

[19] W. Brent Boning. *Metrics for N80: Considerations in Selection and Application to the Programming Process.* CNA Research Memorandum D0005650.A2. Apr. 2002.

[20] James Jondrow, Laura Junor, and Ted Jaditz. *A Sortie Generation Model for DRRS-N.* CNA Research Memorandum 0009204.A2. Dec. 2003.

[21] Laura J. Junor and Peter J. Francis. *Developing Readiness Metrics: An Application of the Coast Guard's Readiness Management System (RMS).* CNA Research Memorandum D0008285.A2. Jan. 2004.

[22] Darlene Stafford and Carol Moore. *A Metric for Surge Capability: Personnel.* CNA Research Memorandum D0009694/A2. Aug. 2004.

[23] Laura J. Junor et al. *Toward a Theory of DoD Readiness Metrics: Initial Thoughts on the Scope and Characteristics of Metrics.* CNA Annotated Briefing, D0006111/A2. Apr. 2002.

[24] Scott Davis, James Jondrow, and Chad Dacus. *Aggregating Readiness Metrics.* CNA Research Memorandum D0017960.A2. May 2008.

[25] Office of the Under Secretary of Defense for Personnel and Readiness (OUSD (P&R)). *Department of Defense Strategic Language List.* Memorandum. Mar. 9, 2011.

[26] "Language Readiness Index Software Manual, Version 4.5." *Defense Readiness Reporting System website,* last accessed Jan. 22, 2012, at http://www.drrs.org.

[27] U.S. Department of Defense Instruction (DODI) 7280.03. *Foreign Language Proficiency Bonus (FLPB).* Aug. 20, 2007.

[28] Peggy A. Golfin et al. *Effectiveness of the Foreign Language Proficiency Bonus (FLPB).* CNA Research Memorandum DRM-2012-U-000734-1REV. Apr. 2012.

This page intentionally left blank.

List of tables

This page intentionally left blank.

CNA
ANALYSIS & SOLUTIONS

4825 Mark Center Drive, Alexandria, VA 22311-1850 703-824-2000 www.cna.org