NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AN ANALYSIS OF FACTORS PREDICTING RETENTION AND LANGUAGE ATROPHY OVER TIME FOR SUCCESSFUL DLI GRADUATES

by

Oleg Green

June 2022

Thesis Advisor: Samuel E. Buttrey
Second Reader: Colby J. Smithmeyer

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2022	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE AN ANALYSIS OF FACTORS PREDICTING RETENTION AND LANGUAGE ATROPHY OVER TIME FOR SUCCESSFUL DLI GRADUATES	**5. FUNDING NUMBERS**
6. AUTHOR(S) Oleg Green	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The Defense Language Institute Foreign Language Center (DLIFLC) is the Department of Defense multi-service school that provides resident instruction in more than a dozen languages to thousands of students annually. Students have to pass the Defense Language Proficiency Test (DLPT) with a score of 2 or better on listening and reading parts of the test to graduate. Service members have to re-test annually after they graduate to maintain their qualifications and additional pay. Some service members maintain their proficiency after graduating better than others, however, and many show a deterioration of proficiency over time and require additional training. DLIFLC needs to better understand how graduates' language skills evolve after they leave the school, to justify future adjustments and enhancements to the program.

Using the data collected by DLIFLC and the Defense Manpower Data Center we developed a logistic regression model to determine what factors are associated with the atrophy of the acquired language skills within the first year after graduation. We also looked at the long-term survival probabilities for DLPT scores using Kaplan-Meyer estimators by stratifying data into subsets. Both methodologies have shown that overall GPA is the most important predictor of the score longevity. Service branch, language category, and initial DLPT scores were shown to be significant discriminators of the test scores' survival over time.

14. SUBJECT TERMS Defense Language Institute Foreign Language Center, DLI, Defense Language Proficiency Test, DLPT, logistic regression	**15. NUMBER OF PAGES** 67
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

AN ANALYSIS OF FACTORS PREDICTING RETENTION AND LANGUAGE ATROPHY OVER TIME FOR SUCCESSFUL DLI GRADUATES

Oleg Green
Major, United States Army
BBA, Campbell University, 2005
MS, Boston University, 2007

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2022

Approved by: Samuel E. Buttrey
Advisor

Colby J. Smithmeyer
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Defense Language Institute Foreign Language Center (DLIFLC) is the Department of Defense multi-service school that provides resident instruction in more than a dozen languages to thousands of students annually. Students have to pass the Defense Language Proficiency Test (DLPT) with a score of 2 or better on listening and reading parts of the test to graduate. Service members have to re-test annually after they graduate to maintain their qualifications and additional pay. Some service members maintain their proficiency after graduating better than others, however, and many show a deterioration of proficiency over time and require additional training. DLIFLC needs to better understand how graduates' language skills evolve after they leave the school, to justify future adjustments and enhancements to the program.

Using the data collected by DLIFLC and the Defense Manpower Data Center we developed a logistic regression model to determine what factors are associated with the atrophy of the acquired language skills within the first year after graduation. We also looked at the long-term survival probabilities for DLPT scores using Kaplan-Meyer estimators by stratifying data into subsets. Both methodologies have shown that overall GPA is the most important predictor of the score longevity. Service branch, language category, and initial DLPT scores were shown to be significant discriminators of the test scores' survival over time.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

CAT	category
CONUS	Continental United States
DLI	Defense Language Institute
DLIFLC	Defense Language Institute Foreign Language Center
DLPT	Defense Language Proficiency Test
DOD	Department of Defense
FY	fiscal year
NPS	Naval Postgraduate School
OCONUS	Outside Continental United States
OPI	Oral Proficiency Interview
ROC	Receiver Operating Characteristic
SDB	student database

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The Defense Language Institute Foreign Language Center (DLIFLC) is the Department of Defense multi-service school that provides resident instruction in more than a dozen languages to thousands of students annually. Students have to pass the Defense Language Proficiency Test (DLPT) with a score of 2 or better on listening and reading parts of the test to graduate. Service members have to re-test annually after they graduate to maintain their qualifications and additional pay. After graduating, some service members can maintain their proficiency better than others. Many service members show a deterioration of proficiency over time and require additional training. DLIFLC needs to better understand how graduates' language skills evolve after they leave the school, to justify cost-effective future adjustments and enhancements to the program.

The purpose of this study was to identify factors associated with the atrophy of acquired language skills for successful DLIFLC graduates over time. The objective was to gain a better understanding and potentially mitigate these atrophy drivers. We used a mixed approach that involved two different techniques to conduct our research. After shaping the problem as a survival analysis, we were able to observe that a significant drop in both reading and listening DLPT scores happens within the first year after graduation, likely during the first annual DLPT retest. The data showed that only 75.5% of the listening scores and 78.2% of reading scores survived past the first year unchanged.

Two separate logistic regression models were fitted onto data to predict whether the listening and reading scores would drop within the first year. For the listening model, the overall GPA in the course is the most important predictor of the score longevity. Students with higher GPAs have better odds of maintaining their listening test scores after graduation. Senior enlisted and officer students also have higher odds of maintaining their scores than enlisted and junior enlisted students. The OCONUS immersion program is a factor that increases the odds of maintaining the score. The surprising finding was that the students who scored higher on the listening portion of the DLPT have lower odds of keeping the same scores for a year after they graduated from DLIFLC. Navy, Air Force,

and Marine students do worse than Army students, and Categories 3 and 4 languages do worse than Categories 1 and 2.

The importance of predictor variables was generally the same for the reading model as for the listening model. The key differences were: immersion of any kind was not statistically significant in this model; being in the Enlisted group (E4–E6) increased the odds of keeping the reading score; and students in the Senior Enlisted and Officer group had even better odds. Just as with the listening model, the reading model shows decreased odds of success for recycled students; however, post-DLPT students have about the same odds of success as the start-to-finish students in the reading model. Unlike the listening model, which shows differences among all services, Air Force students are not statistically different from the Army students, or the base case.

We conducted an additional, long-term survival analysis using the understanding of the factors from our one-year models. To accomplish that we stratified our data into subsets and used Kaplan-Meyer estimators to highlight the differences between the subsets. The long-term survival probabilities in general supported the one-year model findings, with some exceptions. The Army students have shown the best survival rates for both scores in the short-term model, but the Air Force students had better survival probabilities in the long run. The immersion program was not a significant factor in predicting long-term scores survival when controlled for GPA.

This research provides DLIFLC with insights into factors that influence students' ability to retain acquired language skills after they graduate. The findings allow DLIFLC staff to take a closer look at the program to decide how to improve or change it to better meet its goal of teaching language skills that are enduring and long-lasting.

I. INTRODUCTION

This research seeks to identify the specific factors associated with language retention and atrophy over time for successful DLIFLC graduates. DLIFLC requires an understanding of these factors to potentially mitigate the factors that drive atrophy in language skills.

A. MISSION AND GOALS OF DLIFLC

The Defense Language Institute Foreign Language Center (DLIFLC) is the Department of Defense (DOD) multi-service school that provides resident instruction in more than a dozen languages to thousands of students annually. Programs of study vary from 26 to 64 weeks, depending on the difficulty of a language (Defense Language Institute Foreign Language Center 2019). Students can graduate from the program by completing the classes as assigned, or by recycling, re-languaging, taking enhancement training, or a combination of these methods. Passing Defense Language Proficiency Test (DLPT) with a score of 2 or better on the listening and reading parts of the test is a graduation requirement. Not all students graduate in the end, and of those who do graduate, only a fraction stay in the military after their initial enlistment. Service members continue to take DLPT annually, and many show a deterioration of proficiency over time and require additional training.

The language training at DLIFLC is expensive, and costs increase with alterations in the program of instruction. Additional training, such as the immersion program, also adds to the cost. Refer to Appendix A for the estimated costs per student for fiscal year (FY) 2019. DLIFLC needs to better understand what factors influence the language skills retention of service members and how these skills evolve after they leave the school. This understanding will help DLIFLC to justify future adjustments and enhancements to the program to make it more cost-effective.

B. DEFENSE LANGUAGE PROFICIENCY TEST (DLPT)

Defense Language Proficiency Test (DLPT) is the Department of Defense standardized testing system for measuring foreign language proficiency. The DLPT consists of two portions: reading comprehension and listening comprehension. Scores range from 0 to 3, and each level has a + modifier: for example, a test taker can score 1+ on the listening portion, and 2+ on the reading portion of the test. Achieving at least 2 on the listening part and at least 2 on the reading part is a current graduation requirement at DLI. The operational requirements of the intelligence agencies require a higher level of proficiency, and to bridge the gap DLIFLC has been mandated to increase their graduation requirements to a 2+/2+ by the beginning of FY 2023 (Department of the Army 2015). Service members have to re-test annually after they graduate to maintain their qualifications and additional pay.

C. THE PROBLEM

After graduating from DLI some service members can maintain their proficiency better than others, and many show a deterioration of proficiency over time and require additional training. DLI was interested in analyzing the factors that are associated with the atrophy of acquired language skills over time after graduation to gain a better understanding and potentially mitigate the atrophy drivers.

The specific goal is to identify significant predictors that the school can affect internally to improve outcomes following graduation. Developing a meaningful and interpretable prediction model will help DLIFLC to better assess the return on investment of various training techniques and modules.

D. THESIS ORGANIZATION

This thesis is organized into five chapters that explore and analyze the factors associated with the survival of language training that students receive at DLIFLC. Chapter II is a literature and methodology review that looks at past DLIFLC studies and describes the methodology used in this study. Chapter III describes the data used in this study and addresses how the survival data used as the basis for the analysis was created. Chapter IV

explains the analysis performed and the analytic results. Chapter V contains the conclusions and recommendations that the analysis produced.

THIS PAGE INTENTIONALLY LEFT BLANK

II. LITERATURE REVIEW AND METHODOLOGY

A. LITERATURE REVIEW

In this section, we take a look at the past work that is relevant to this study.

1. Previous Work

Over the past few years, several Naval Postgraduate School graduates studied factors that affect the attrition and success of DLIFLC students. In chronological order:

An Analysis of Factors Predicting Graduation of Students at Defense Language Institute Foreign Language Center (Wong 2004). The study is based on DLIFLC internally collected data from 1998 to 2003. The author used four logistic regression models separated by language category (I, III, IV, and all languages together) to model factors that predict successful graduation. The analysis showed the importance of DLAB scores, recycle/re-language status, service affiliation, and gender for graduation success.

A Statistical Analysis of Individual Success After Successful Completion of Defense Language Institute Foreign Language Center Training (Hinson 2005). The author used DLIFLC data for students from 1997 to 2000 and combined it with DMDC data for the same students showing the lengths of service, loss dates, and Foreign Language Proficiency Pay (FLPP) to determine students' success or failure after graduation. Hinson used a classification tree and logistic regression model to determine that service affiliation, contract lengths, and gender were primary predictors of a successful outcome.

Analysis of Korean Academic Attrition at the Defense Language Institute Foreign Language Center (Haupt 2014). Haupt used DLIFLC student database records for students who studied the Korean language from 2006 to 2013. He used a logistic regression model to predict failure to meet graduation requirements in the course. He determined that significant factors that contribute to the risk of failure were recycled status, service affiliation, and semester great point averages.

Student success factors at Defense Language Institute Foreign Language Center (Bermudez-Mendez 2020). Bermudez used DLIFLC student database records for all

students from 2011 to 2018 to determine factors that contribute to students meeting L2+/R2+ standard on DLPT. He also studied the effects of the immersion program on student success. The author used logistic regression to model the data and concluded that performance in advanced language classes was the most significant factor in predicting success. Other significant factors were the DLAB score, prior language experience, and language category. The immersion program did not significantly contribute to improvements in DLPT scores, after accounting for selection bias.

Student Achievement Indicators at Defense Language Institute Foreign Language Center (Brenner 2021). The author used DLIFLC student records from 2011 to 2018 to determine factors contributing to students achieving the L2+/R2+ standard on DLPT. Brenner used random forest and neural network models for his research. He identified that language background, prior language source, prior language experience, and prior language proficiency were statistically significantly related to student success.

Our research used many of the same variables to model the data. In the interest of full disclosure, the analysis in this thesis used ideas and techniques from all of these projects to transform data for this thesis.

2. Air Force Language Retention Study 2013

In *Modeling Second Language Change Using Skill Retention Theory*, Shearer (2013) used skill retention theory to explain the acquisition, retention, attrition, and reacquisition of a foreign language using DLIFLC student data. The skill retention theory uses three stages of learning to explain how skills decay over time based on the level of acquired proficiency (Kim et al. 2011). Shearer used Survival Analysis in his work to identify a relationship between a linguist's proficiency level at graduation and the length of time after graduation until a change in proficiency occurs. The author also looked at the effects of additional training on rates of reacquiring language skills after a period of atrophy. He used internally collected DLIFLC student data and Defense Manpower Data Center (DMDC) data for DLPT scores after graduation.

Shearer's primary conclusions were: second language skills are similar to sensory-motor and other cognitive skills, and skills retention theory represents language change

well and can be used as a general language change theory. Skills retention theory connects the deterioration of a given skill with the stages of a learning process—higher mastery of skill leads to a slower rate of its atrophy (Kim et al. 2011). He emphasized the role of the additional follow-on courses to increase language retention and determined that the higher proficiency levels at graduation, expressed by the DLPT scores, would slow down the decay rates. Figure 1, adapted from the dissertation, shows survival curves for DLPT scores, listening on the left, and reading on the right. The proportion of unchanged (surviving) scores is plotted against the time that passed since students graduated DLI. Survival curves for students who scored 3 at graduation are depicted in blue, those who scored 2+ have red curves, and those who scored 2 have survival curves depicted in black. The decay in proficiency down to 1+ level takes longer on average for students who scored 2+ on listening or reading portions of DLPT at graduation and takes even longer for those that scored 3. In other words, students who scored higher on DLPT at graduation retain their scores longer than those who scored lower. Shearer measured the survival time as the time that passed from the graduation until the score dropped down to 1+.

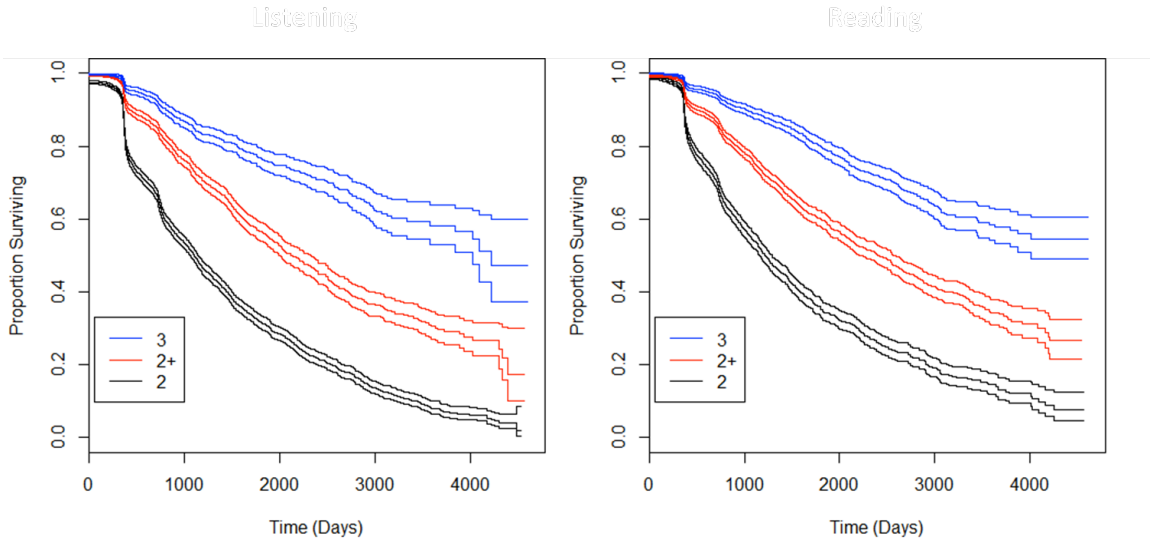


Figure 1. Length of Time since Graduation to 1+ with 95% Confidence Intervals. Source: Shearer (2013).

B. ANALYSIS METHODOLOGY

This section describes the analytical methods and techniques used in the thesis.

1. Logistic Regression

The first approach to analyze our data was to use logistic regression, the model that predicts a binary outcome. The model estimates the probability that an instance of data belongs to one of two classes, and based on probabilities assigns it a label. Géron (2019) states that the logistic regression model calculates a weighted sum of the input features (adding a bias term), but instead of producing a direct result like the linear regression, logistic regression yields the logit of the result, which is passed to a sigmoidal function (σ) that produces a number between 0 and 1, that is

$$\hat{p} = \sigma(x^T \theta) \tag{1}$$

and

$$\sigma(t) = \frac{1}{1+e^{-t}}, \tag{2}$$

where θ is a vector of parameters.

The dataset used in the research was split into training and test sets (80% and 20% of all data, respectively) using a random selection of data points. Only the main effects of the full set of predictor variables were used to fit the model on training data. The goal was to produce a model that was parsimonious and easily interpretable, even at the cost of some accuracy. The R package `glm` was used to fit the model (R Core Team 2019).

We used a stepwise selection algorithm to identify variables to include in the model based on their importance. According to James et al. (2013), the stepwise selection algorithm starts with the full set of predictor variables and removes or adds them one at a time iteratively while recalculating the Akaike information criterion (AIC) at each step. The AIC is defined by a penalized maximum likelihood. The process goes on until the model has only the combination of the most significant predictors with the smallest AIC left.

To assess the performance of our logistic regression model, we created a confusion matrix, also called a classification table, that shows prediction results on the test data. An example of the confusion matrix, adapted from James et al. (2013) is shown in Figure 2.

	Predicted Positive	**Predicted Negative**
Positive	True Positive (TP)	False Negative (FN) Type I error
Negative	False Positive (FP) Type II error	True Negative (TN)

Figure 2. Generalized Confusion Matrix for a Binary Classifier

We draw several model assessment metrics from the confusion matrix. The correct positive and negative predictions are shown on the main diagonal of the matrix. The overall accuracy of the model is calculated as (TP + TN)/Total Population. The accuracy shows the proportion of correct predictions from the total number of predictions.

The sensitivity of the model, also called recall, is another important metric that shows the ratio of positive instances that our model correctly detected (Géron, 2019). It is calculated as TP/(TP + FN).

Another metric is the specificity of the model which shows the proportion of negative instances that our model correctly identifies (James et al. 2013). Specificity is calculated as TN/(TN + FP).

The precision or Positive Predictive Value (PPV) shows the accuracy of the positive predictions (Géron 2019), and is calculated as TP/(TP + FP). Negative Predictive Value (NPV) shows the accuracy of negative predictions, and it is calculated as TN/(TN + FN).

We also look at the Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC) to assess how well our model performs. The ROC simultaneously displays two types of errors for all possible thresholds (James et al. 2013). It is a plot of

sensitivity against 1−specificity. The AUC summarizes the model over all possible thresholds and shows the overall performance of the classifier (James et al. 2013).

Each variable in our final model has an estimated odds ratio. The odds ratio shows the odds of a positive outcome for each level associated with categorical variables, compared to the baseline; and the odds associated with an increase of one unit for continuous variables (James et al. 2013). Variable importance and the odds ratio for each variable were used to interpret the model.

2. Kaplan-Meier Estimator (KME)

The second approach used was Survival Analysis. A study that deals with time until failure or time-to-event (often used in clinical studies) is commonly referred to as Survival Analysis. If we represent the lifetimes of the observations by a distribution T, then the expected proportion of the observations for which the event of interest has not yet occurred by time t can be expressed by a function

$$S(t) = P_r(T > t). \tag{3}$$

S(t) is a survival function, and it represents the probability that for the randomly selected individual from the data set the event of interest will happen after time t. The survival function is a non-decreasing function with $S(0) = 1$ (Aalen and Gjessing 2008).

The main challenge with analyzing survival data is the presence of incomplete observations. By the time the study ends, the event of interest will never have happened for some subjects of the study. In an example of clinical studies, some patients have died at a time t_i, and some are still alive at the end of the study or dropped out of the study altogether before the end. The real survival time is unknown for the last two groups. These survival times are called right-censored. The mix of known and right-censored survival times cannot be handled by the usual statistical methods. Even calculating a simple mean does not make sense for this data, which makes finding a standard deviation, or performing a t-test impossible (Aalen and Gjessing 2008). Ignoring right-censored values or capping them at their known values would introduce bias into the calculations—we would be ignoring longer (possibly much longer) survival times in our study.

The Kapan-Meier Estimator (KME) is the non-parametric method of choice for estimating the survival function for the data where censored observations are present. The KME combines a set of conditional probabilities like

Probability (survive past t_i | survive until t_i) =
Probability (survive past t_i) / Probability (survive until t_i),
estimated by (# surviving past t_i) / (# surviving to t_i).

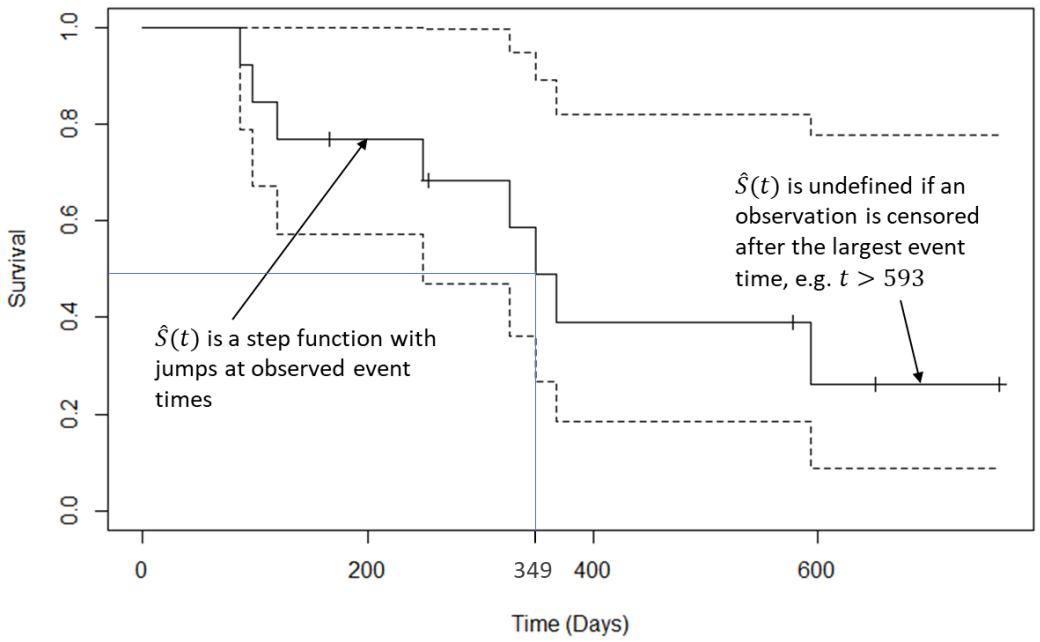
For each lifetime that ends in death or censoring, the (# surviving to t_i), called number at risk, decreases. The final KM estimate is the product of those conditional probabilities (Aalen and Gjessing 2008). We use a small numerical example to demonstrate the works of the estimator (see Table 1).

Table 1. Kaplan-Meyer Estimator Example

Times t_i, ordered	**Censored, 1–yes, 0–no**	**# surviving past t_i**	**# surviving to t_i**	**# surviving past t_i / # surviving to t_i**	**Product of probabilities**
87	0	12	13	12/13	**0.923**
98	0	11	12	11/12	**0.846**
120	0	10	11	10/11	**0.769**
166	1	10	10	10/10	0.769
249	0	8	9	8/9	**0.684**
254	1	8	8	8/8	0.684
326	0	6	7	6/7	**0.586**
349	0	5	6	5/6	**0.488**
367	0	4	5	4/5	**0.391**
578	1	4	4	4/4	0.391
593	0	2	3	2/3	**0.260**
651	1	2	2	2/2	0.260
761	1	1	1	1/1	0.260

1. The KM estimate only jumps at event times (**bold**)
2. S(t) is undefined after the last recorder event (at 593)

The convenient visual representation of the KM estimation is a KM curve, as shown in Figure 3:



Notes: Censoring is indicated by the vertical tic marks; dotted lines show 95% CI

Figure 3. Kaplan-Meier Curve Generated by R Studio from the Sample Data in Table 1.

KME curves are very informative. The length of the horizontal segments represents the survival duration for the interval (along X-axis). The vertical down steps represent the change in the survival function as the estimate progresses (Rich et al. 2008).

Once KM estimation is complete, the median survival time can be calculated as the first event time after the survival probability drops below 0.5. In the example above, it is 349 days. The median is a preferred measure of central tendency for survival analysis, due to the frequent skewness of data. There are various ways to construct confidence intervals for the median survival time and to estimate the variance of a survival function. (Aalen and Gjessing 2008). These methods are outside of the scope of this thesis.

The common approach to using KM estimates in the analysis is to stratify the data using one or more covariates and compare the curves. To quantify the differences between

curves to assess statistical significance, the log-rank test is normally used. The log-rank test calculates the chi-square (χ^2) for each event time for each group and sums the results. The summed results for each group are added to derive the ultimate chi-square to compare the full curves of each group (Rich et al. 2008).

There are several considerations with using survival analysis. First, censoring for survival data must be non-informative. Non-informative censoring means that it does not contain any information on the subject in the dataset, other than the censoring time itself (Aalen and Gjessing 2008). Second, analysts have to choose meaningful ways to stratify data by covariates, which can be challenging. The continuous variables have to be grouped into categories. Finally, as variance gets larger as time progresses since the number of survivors gets smaller, making estimations at the far right of the curve less precise (Rich et al. 2008).

THIS PAGE INTENTIONALLY LEFT BLANK

III. DATA DESCRIPTION

A. DATA PREPARATION

Two separate data sets were used in this study, each provided by the DLIFLC's Academic Data Systems department. The department maintains the student database, which contains records describing demographic and academic performance data for students from admission to graduation (or dis-enrollment). The subset of historical data used contained records for students ranging from FY 2010 to FY 2018. It had a total of 27,714 records with 53 variables describing each record. Data for the student database is collected from the enrollment records, questionnaires, class grades, and test scores. It has a substantial amount of missing and erroneous data.

The majority of the students have a single record in the database. This represents a case where a student was enrolled in school and graduated, or was disenrolled and never restarted the study. Some students have more than one record, for various reasons. The first reason is if the student studied one language, graduated, and came back some years later and studied a different language. The second reason, recycling or re-languaging, is if the student started the study, was not successful, and had to restart at an earlier point in the same language program or was reassigned to a different, usually easier, language. The third reason, the post-DLPT program, is when students complete the program but do not score high enough on the DLPT to graduate, they are enrolled in additional training (normally eight weeks) to prepare for a re-test. For recycled, re-languaged, and post-DLPT students grades and demographic information is split between different records.

Records of interest from this first dataset included only records for students who have successfully graduated from DLIFLC and had valid DLPT scores at graduation. We removed all records for USCG and civilian students since they were not in the population of interest. The resulting intermediate dataset still had multiple records for students who successfully studied different languages and graduated.

The second dataset was created by the Defense Manpower Data Center (DMDC) and contained DLPT test scores of DLIFLC graduates from FY 2011 to FY 2022 (partial,

up to December 2021). It contained 101,858 records, each with 11 variables, such as reading, listening, and speaking scores, dates, and test language names. Approximately 16,000 unique students had records in the data set, ranging from one to about 25 tests for each student. Multiple students had recorded test scores for more than one language: for example, some students who studied Russian also took tests in Ukrainian and Bulgarian. A majority of Oral Proficiency Interview (OPI) scores were missing or had invalid scores and were not used in this study.

Records of interest from the second dataset included test scores for students present in the first intermediate dataset who had valid test scores for the languages that they studied at DLIFLC. The two datasets were merged using masked student id numbers and language names that students learned at DLIFLC, creating a combined dataset for the analysis. Records for students in the intermediate dataset who did not have recorded tests in the DMDC test set were excluded. The final dataset contained 11,521 records and had 58 variables.

B. RESPONSE VARIABLES

DLPT scores and dates from the DMDC dataset were not used directly in the final dataset used for analysis. Instead, six response variables were created out of these records. The first response variable, `ldays` is a numeric value that shows the number of days that passed after graduation from DLIFLC until a student's listening DLPT score has dropped by any number of points. For example, if a student scored 2+ on the listening portion of DLPT at graduation, then scored 2+ on his test 350 days later (~ one year mark), and scored 2 on his test 762 days after graduation (~ two years mark), his value for `ldays` would be 762. His subsequent scores after the first drop in the score are not taken into account, even if he scored 2+ or higher sometime later. The time to the first score drop is the value of the interest in our research. If a student's score never drops on record, the number of days from graduation until the last available test date is recorded, the maximum known time without a drop in the listening score.

The second response variable is `l.chng`. It is a binary variable that denotes if the drop in the listening score happened (1 = yes, 0 = no). In other words, it shows if the `ldays`

value is right-censored or not, 0 or 1, respectively. In the above example, a student that has `ldays` = 762 would have `l.chng` = 1, meaning the event of interest, the drop in the score was observed, and the value of `ldays` is not right-censored. If a student's listening score never dropped, and his maximum observed time without a drop is recorded, `l.chng` is set to 0, showing the right-censoring.

The third response variable is `ldrop`, also a binary value showing if a student's listening score dropped within 390 days after graduating from DLIFLC (~ one year). The value is set at 0 if the score dropped, and 1 if the score has not dropped within the first year following graduation. Going back to the same example, `ldrop` will be set as 1, since the score drop occurred at 762 days, which is greater than 390. If the `ldays` value was less than 390, and `l.chng` was 0, meaning that we only have scores that fall within the first year after graduation, and the score never went down, `ldrop` was set as 1, the score has not dropped.

The last three response variables: `rdays`, `r.chng`, and `rdrop` are exactly the same as the first three described, but apply to the drop in the reading scores. Variables `ldays`, `l.chng`, `rdays`, and `r.chng` were created for use in survival analysis, and variables `ldrop`, `rdrop` were created for logistic regression modeling (Table 2).

Table 2. Response Variables

Name	Symbol	Classification	Description
Listening, days to drop	ldays	Continuous	Number of days until listening score drops
Listening, event	l.chng	Categorical	0 (right-censored value) 1 (event happened)
Listening, 1-year	ldrop	Categorical	Listening score dropped within one year 0 (Yes), 1 (No)
Reading, days to drop	rdays	Continuous	Number of days until reading score drops
Reading, event	r.chng	Categorical	0 (right-censored value) 1 (event happened)
Reading, 1-year	rdrop	Categorical	Reading score dropped within one year 0 (Yes), 1 (No)

C. PREDICTOR VARIABLES

Many available variables, such as class dates and class numbers, were not included in the model, such as class dates and class numbers. Some variables were transformed for use in our analysis. Student's ranks were grouped to create a new variable Rank_Group, where E1–E3 were grouped as Juinor_Enlisted, E4–E6 as Enlisted, and E7–E9 and officers as Senior_E_Officer. Class grades were grouped into Language and Culture classes, and letter grades were converted to numeric GPAs. The letter-to-grade conversion chart is in Appendix B (adapted from Bermudez-Mendez 2020). The list of variables that were considered in the model is described in Table 3.

Table 3. Predictor Variables

Name	Symbol	Classification	Description
Service Branch	Svc	Categorical	USA (Army) USN (Navy) USMC (Marine Corps) USAF (Air Force)
Language Category	Lang.Cat	Categorical	Difficulty of Language: 1 (CAT I) 2 (CAT II) 3 (CAT III) 4 (CAT IV)
DLAB	DLAB	Continuous	Scores from 71 to 159
DLAB Waiver	DLAB.Waiver	Categorical	Y (Yes) N (No)
Gender	Gender	Categorical	M (Male) F (Female)
Rank Group	Rank_Group	Categorical	Junior Enlisted (E-1, E-2, E-3) Enlisted (E-4, E-5. E-6) Sr. Enlisted & Officers (E-7 and above)
Input Status	In_Status	Categorical	I (New Input) J (Relanguaged) P (Post-DLPT) Q (Recycle – Same Course)

Name	Symbol	Classification	Description
Overall GPA	GPA	Continuous	From 1.8 to 4.0
Listening Score (at Graduation)	DLPT.L	Continuous	From 6 to 30
Reading Score (at Graduation)	DLPT.R	Continuous	From 6 to 30
Listening Score (at Graduation)	DLPT_L	Categorical	16 (L1+ and below) 20 (L2) 26 (L2+) 30 (L3)
Reading Score (at Graduation)	DLPT_R	Categorical	16 (R1+ and below) 20 (R2) 26 (R2+) 30 (R3)
Immersion	Immersion	Categorical	O (OCONUS) C (CONUS) U (Unknown Location) N (No immersion)
Elementary Language Group	FL1XX_Lang_Classes	Continuous	Average grade in FL101, FL102 and FL110
Intermediate Language Group	FL2XX_Lang_Classes	Continuous	Average grade in FL201, FL202 and FL210
Advanced Language Group.	FL3XX_Lang_Classes	Continuous	Average grade in FL301, FL302 and FL310
Elementary Culture Group	FL1XX_Culture_Classes	Continuous	Average grade in FL120 and FL140
Intermediate Culture Group	FL2XX_Culture_Classes	Continuous	Average grade in FL220 and FL240
Advanced Culture Group	FL3XX_Culture_Classes	Categorical	Average grade in FL320 and FL3240

D. DESCRIPTIVE STATISTICS

The resulting dataset used in this study contained 11,521 records. In this section, we present statistics that help us better understand the data.

1. Students with Multiple Observations

There are 156 students in our dataset who have more than one record. Of those, 153 have two records each, and two students have three records each. These students have

successfully enrolled and graduated from DLIFLC more than once in the period from 2011 to 2018. The remaining 11,227 students have single records in our dataset.

2. Distribution of Students by Service Component and by Rank

Figure 4 shows the distribution of students in our dataset by service component. Air Force students make up 39.1% (4511) of the total, the largest group in the set. Army students account for 36.5% (4207) of records, Navy for 13.5% (1559), and Marines have the smallest proportion of students at 10.9% (1262).

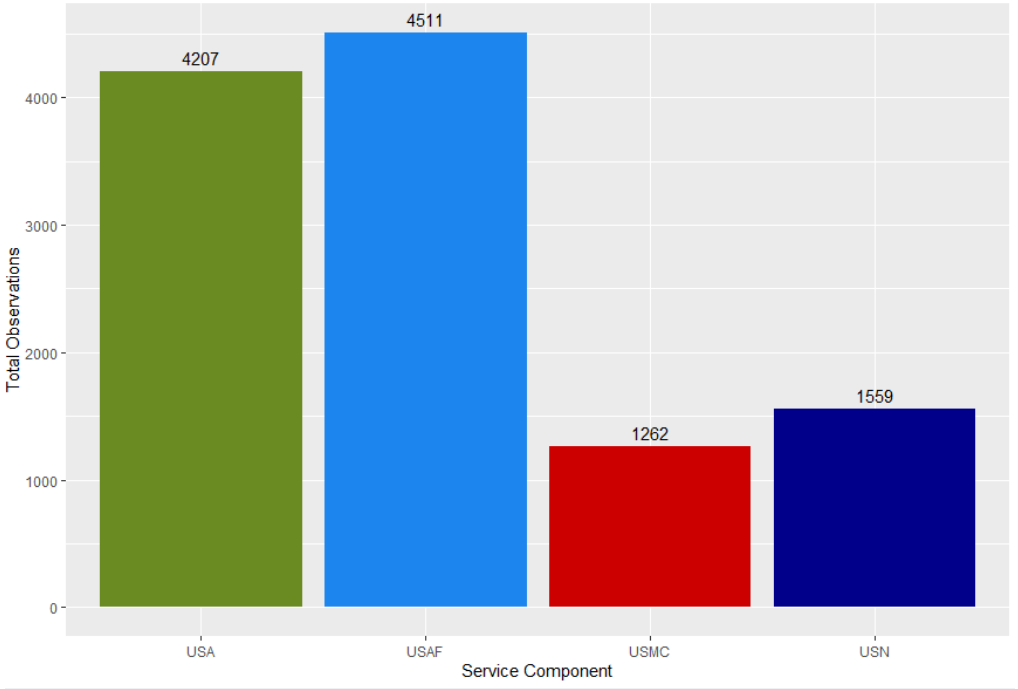


Figure 4. Number of Observations by Service Component

Records of the junior enlisted service members (E1–E3) make up the largest proportion of the dataset at 62.9% (7257), followed by enlisted (E4–E6) at 26.7% (3087), and then senior enlisted and officers (E7 and above) at 10.4% (1195).

3. Distribution of Observations by Language and Language Categories

In our dataset, we have records for students that studied 27 different languages, with the number of observations ranging from three for Punjabi (PJ) to 1818 for

Arabic-Modern (AD). Records for eight languages: Arabic-Modern (1818), Persian-Farsi (1616), Chinese-Mandarin (1599), Korean (1111), Russian (1071), Spanish (1043), Pushtu-Afghan (971), and French (604) make up 85% of the data set. For the full list of languages along with their two and three-letter codes refer to Appendix B, List of DLIFLC languages.

The majority of observations, 52.6% in the dataset are for students who studied Category 4 (the hardest) languages, followed by Category 3 at 30.6%, with Category 1 and 2 languages comprising 15.0% and 1.7% of all observations. Figure 5 shows the distribution of languages in the dataset by category.

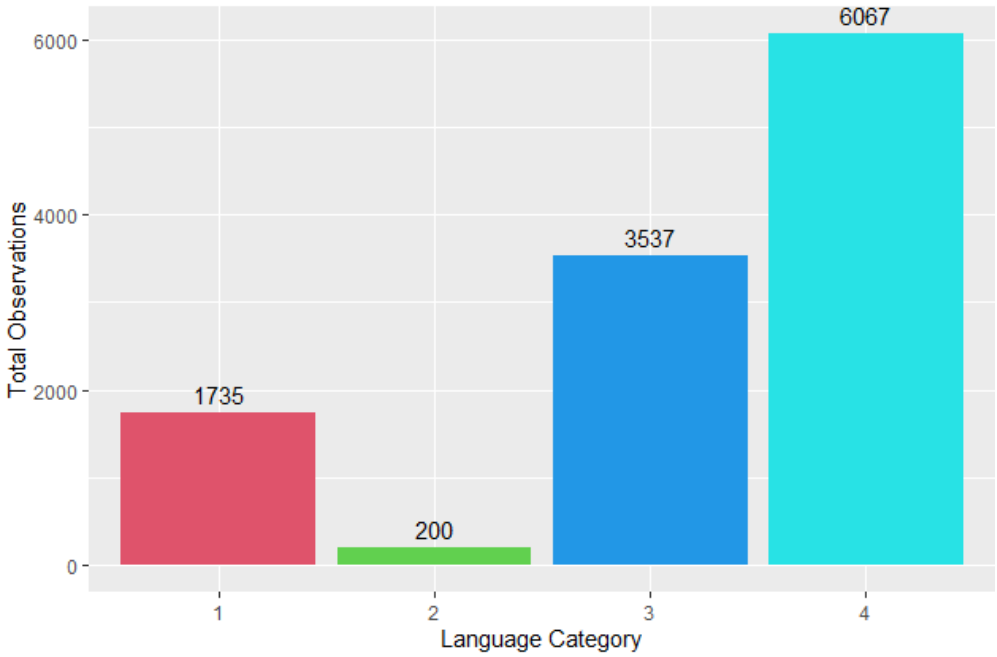


Figure 5. Number of Observations by Language Category

4. Other Statistics of Interest

Our dataset has more observations for male students (8702) than for female students (2837), 75.4% and 24.6%, respectively.

There are 8917 observations or 77.3% of the total dataset belonging to students who have not attended any immersion training programs. A total of 1423 students (12.3%)

attended the OCONUS immersion program, and 287, or 2.5% of the total attended CONUS immersion.

Of all records in the data set, 84.0% (9697) of all observations are for students who enrolled and graduated from the program in a single go. For the rest, 7.4% of all students attended post-DLPT training (850 observations), 7.3% were recycled at least once before graduating (837 observations), and 1.3% of students in our dataset were re-languaged before graduating successfully (155 observations).

5. Survival Data

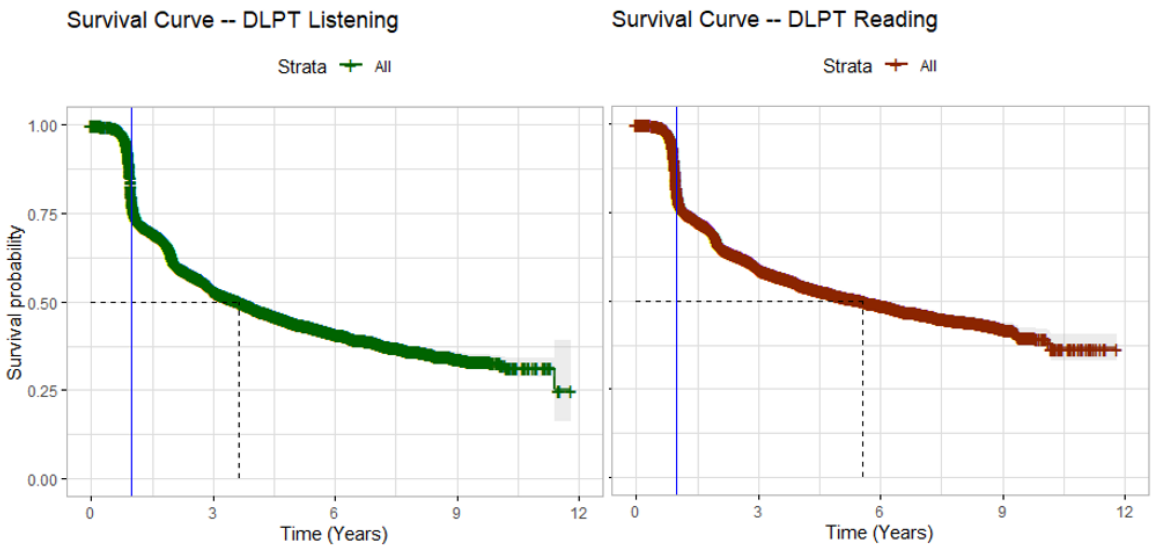
About 24.5% of all students experienced a drop in DLPT listening score before or at the one-year mark after graduation, while 21.8 % experienced a drop in DLPT reading score before or at the one-year mark in our data set.

IV. ANALYSIS AND RESULTS

In this chapter, we explain the steps we took in our analysis, results, and interpretation.

A. EXPLORATORY ANALYSIS

Service members must take a DLPT annually (with some exceptions) to maintain their qualifications for the job and continue to receive their language incentive pay. After a preliminary analysis, it was obvious that after graduating from DLIFLC students' DLPT scores gradually decrease over time for the cohort, and the most dramatic change happens around the one-year mark. Figure 6 shows the noted drop in listening (on the left) and reading (on the right) DLPT scores. The survival curves show that only 75.5% of the cohort survived the first annual re-test for listening scores and 78.2% for reading scores (where "survival" indicates no downward changes in the score).



Note: Blue vertical line shows one year mark.

Figure 6. Survival Curves for Listening (left) and Reading (right) for the Cohort.

Median survival times for listening and reading scores were different for the cohort; it took 3.65 years for 50% of all listening scores to drop, and 5.56 years for reading scores. That means that listening scores atrophy at a faster rate. There is a possibility that different predictors affect the two rates of atrophy in dissimilar ways.

Based on these conclusions we decided to create two separate response variables for modeling our dataset using logistic regression. Therefore, we have two binary response variables `ldrop` and `rdrop` that denote if the listening or reading score for the observation went down during the first year after graduation. And we then created two separate models to gain insight into the differences between listening and reading score behavior.

B. ONE-YEAR CHANGE MODELS

1. Goodness-of-Fit and Performance

We assess the discrimination power of our models first. Models with an area under the ROC curve between 0.7 and 0.8 are considered to have acceptable discrimination (Hosmer et al. 2013). Both of our models have acceptable discriminating power, with the reading model performing slightly better. See Figure 6 for the ROC plots.

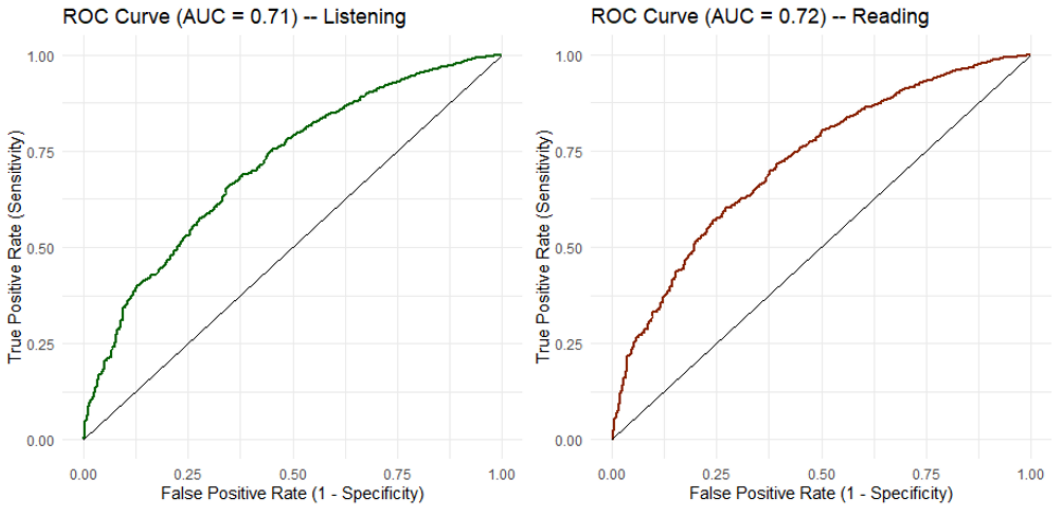


Figure 7. Plots of ROC Curves for Listening (left) and Reading (right) Models.

Next, we take a look at the predictive power of our models. For the listening model, the classification table is presented in Table 4.

Table 4. Listening Model Classification Table

	Score Dropped	**Score Did Not Drop**
Predicted Drop	62	507
Predicted No Drop	40	1699

From the classification table the following measures were derived:

- Accuracy: 0.76, 95% CI : (0.7451, 0.7802)
- Sensitivity: 0.61
- Specificity: 0.77
- Positive Predictive Value: 0.11
- Negative Predictive Value: 0.98

A 0.76 accuracy means that the model correctly predicts score dropping or not 76% of the time. A 0.61 sensitivity means that the model correctly predicts score not dropping for the student whose score has not dropped about 61% of the time. A specificity of 0.77 means that our model can predict score drop for students whose score has decreased about 77% of the time. A positive predictive value of 0.11 means that when we predict that a student's score will not drop, the prediction is correct about 11% of the time. And finally, a negative predictive value of 0.98 means that when the model predicts a drop in the score, the prediction is correct about 98% of the time. The model has a much better predictive power for students whose scores drop, and this is the outcome we want from our model.

For the reading model, the classification table is presented in Table 5.

Table 5. Reading Model Classification Table

	Score Dropped	**Score Did Not Drop**
Predicted Drop	56	440
Predicted No Drop	41	1771

From the classification table the following measures were derived:

- Accuracy: 0.79, 95% CI : (0.7744, 0.808)
- Sensitivity: 0.58
- Specificity: 0.80
- Positive Predictive Value: 0.11
- Negative Predictive Value: 0.98

The reading model is slightly more accurate than the listening model, with all the measures of the predictive power in line with the listening model.

2. Variable Interpretation

The summary of the coefficient estimates is provided in Table 6. Statistically significant predictors are marked in bold. The stepwise logistic regression reduced the number of variables in the model, dropping the ones that were not significantly contributing to the model. The predictors with *p*-values greater than 0.05 left in the model are levels of the categorical predictors.

Table 6. Estimates of Predictor Variables for Listening Model

Factor	Estimate	Std. Error	z value	P-value
(Intercept)	−1.625	0.3418	−4.75	0.000
SvcUSAF	**−0.196**	**0.0645**	**−3.04**	**0.002**
SvcUSMC	**−0.269**	**0.0940**	**−2.86**	**0.004**
SvcUSN	**−0.390**	**0.0846**	**−4.61**	**0.000**
Lang.Cat2	−0.183	0.2256	−0.81	0.417
Lang.Cat3	**−0.489**	**0.0937**	**−5.22**	**0.000**
Lang.Cat4	**−0.444**	**0.0881**	**−5.04**	**0.000**
In.StatusJ	0.005	0.2435	0.02	0.984
In.StatusP	**−0.324**	**0.0977**	**−3.32**	**0.001**
In.StatusQ	**−0.339**	**0.0987**	**−3.44**	**0.001**
Rank_GroupEnlisted	0.007	0.0638	0.11	0.911
Rank_GroupSenior_E_Officer	**0.667**	**0.1183**	**5.64**	**0.000**
GPA	**1.593**	**0.0976**	**16.33**	**0.000**
ImmersionC	0.170	0.1661	1.02	0.307
ImmersionO	**0.255**	**0.0866**	**2.94**	**0.003**
ImmersionU	−0.073	0.0988	−0.74	0.458
DLPT_L20	**−1.503**	**0.1799**	**−8.35**	**0.000**
DLPT_L26	**−2.645**	**0.1830**	**−14.46**	**0.000**
DLPT_L30	**−3.001**	**0.1916**	**−15.66**	**0.000**

Table 7 shows variables in our listening model with their importance, odds ratio, and confidence interval. The odds ratios can be interpreted as increased odds of keeping the same score if a student is exposed to a factor for categorical variables and increased odds of keeping the score for the unit of increase for continuous variables. Variables are sorted in decreasing order of importance, and variables that increase the odds of maintaining a score, where odds ratios are greater than one, are marked in bold. Baseline values for each categorical variable are omitted in the table. They are: DLPT_L16, Rank_GroupJunior_Enlisted, Lang.Cat1, SvcUSA, In.StatusI, and ImmersionN. Variables at the bottom, marked in italics, are statistically insignificant and do not contribute to the model. Variable importance was calculated as the absolute value of the coefficients in the model, using `varImp` function in R `caret` package.

Table 7. Listening Model Interpretation Information

Variable	Importance	Odds Ratio	Lower 95%	Upper 95%
GPA	**16.328**	**4.921**	**4.066**	**5.961**
DLPT_L30	15.659	0.050	0.034	0.072
DLPT_L26	14.457	0.071	0.049	0.100
DLPT_L20	8.354	0.223	0.154	0.312
Rank_GroupSenior_E_Officer	**5.639**	**1.949**	**1.552**	**2.469**
Lang.Cat3	5.222	0.613	0.509	0.736
Lang.Cat4	5.038	0.642	0.539	0.761
SvcUSN	4.607	0.677	0.574	0.800
In.StatusQ	3.440	0.712	0.588	0.865
In.StatusP	3.321	0.723	0.598	0.877
SvcUSAF	3.040	0.822	0.724	0.933
ImmersionO	**2.942**	**1.290**	**1.091**	**1.532**
SvcUSMC	2.858	0.764	0.636	0.920
ImmersionC	*1.021*	*1.185*	*0.861*	*1.654*
Lang.Cat2	*0.811*	*0.833*	*0.542*	*1.315*
ImmersionU	*0.742*	*0.929*	*0.767*	*1.130*
Rank_GroupEnlisted	*0.111*	*1.007*	*0.889*	*1.142*
In.StatusJ	*0.019*	*1.005*	*0.634*	*1.654*

From the listening model, we can tell that the overall GPA in the course is the most important predictor of the score longevity. Students with higher GPAs have better odds of maintaining their listening test scores after graduation. Senior enlisted and officer students also have higher odds of maintaining their scores than enlisted and junior enlisted students. The OCONUS immersion program is a factor that increases the odds of maintaining the score. Students that scored higher on the listening portion of the DLPT, have lower odds of keeping the same scores for a year after they graduated from DLIFLC. Navy, Air Force, and Marine students do worse than Army students, and Categories 3 and 4 languages do worse than Categories 1 and 2.

Estimates of predictor variables and their interpretation for the reading model are described in Table 8 and Table 9.

Table 8. Estimates of Predictor Variables for Reading Model

Factor	Estimate	Std. Error	z value	P-value
(Intercept)	−1.644	0.4197	−3.92	0.000
SvcUSAF	−0.162	0.0684	−2.36	0.018
SvcUSMC	**−0.382**	**0.0964**	**−3.97**	**0.000**
SvcUSN	**−0.520**	**0.0877**	**−5.93**	**0.000**
Lang.Cat2	−0.025	0.2267	−0.11	0.912
Lang.Cat3	**−0.547**	**0.0981**	**−5.58**	**0.000**
Lang.Cat4	**−0.311**	**0.0911**	**−3.42**	**0.001**
In.StatusJ	−0.105	0.2470	−0.42	0.672
In.StatusP	0.029	0.1081	0.27	0.789
In.StatusQ	**−0.436**	**0.1017**	**−4.29**	**0.000**
Rank_GroupEnlisted	**0.212**	**0.0682**	**3.11**	**0.002**
Rank_GroupSenior_E_Officer	**0.876**	**0.1214**	**7.22**	**0.000**
GPA	**1.802**	**0.1025**	**17.57**	**0.000**
ImmersionC	0.416	0.1842	2.26	0.024
ImmersionO	0.217	0.0898	2.42	0.016
ImmersionU	−0.212	0.1011	−2.10	0.036
DLPT_R20	**−1.542**	**0.3057**	**−5.05**	**0.000**
DLPT_R26	**−3.245**	**0.3058**	**−10.61**	**0.000**
DLPT_R30	**−3.803**	**0.3123**	**−12.18**	**0.000**

The importance of predictor variables was generally the same for the reading model as for the listening model. The key differences were: immersion of any kind was not statistically significant in this model, being in the Enlisted group (E4–E6) increased the odds of keeping the reading score, and students in the Senior Enlisted and Officer group had even better odds. Just as with the listening model, the reading model shows decreased odds of success for recycled students; however, post-DLPT students have about the same odds of success as the start-to-finish students in the reading model. Unlike the listening model, which shows differences among all services, Air Force students are not statistically different from the Army students, or the base case.

Table 9. Reading Model Interpretation Information

Variable	Importance	Odds Ratio	Lower 95%	Upper 95%
GPA	**17.569**	**6.059**	**4.959**	**7.413**
DLPT_R30	12.177	0.022	0.011	0.039
DLPT_R26	10.613	0.039	0.020	0.068
Rank_GroupSenior_E_Officer	**7.218**	**2.402**	**1.902**	**3.063**
SvcUSN	5.926	0.595	0.501	0.707
Lang.Cat3	5.581	0.579	0.477	0.700
DLPT_R20	5.046	0.214	0.111	0.373
In.StatusQ	4.287	0.647	0.531	0.791
SvcUSMC	3.965	0.682	0.565	0.825
Lang.Cat4	3.417	0.732	0.612	0.874
Rank_GroupEnlisted	**3.110**	**1.236**	**1.082**	**1.414**
ImmersionO	*2.418*	*1.243*	*1.044*	*1.484*
SvcUSAF	*2.361*	*0.851*	*0.744*	*0.973*
ImmersionC	*2.257*	*1.515*	*1.067*	*2.201*
ImmersionU	*2.098*	*0.809*	*0.665*	*0.988*
In.StatusJ	*0.424*	*0.901*	*0.564*	*1.490*
In.StatusP	*0.268*	*1.029*	*0.835*	*1.275*
Lang.Cat2	*0.110*	*0.975*	*0.633*	*1.543*

C. SURVIVAL ANALYSIS

After we modeled our data using logistic regression, we gained a better understanding of what factors are associated with listening and reading scores atrophy one-year after a student graduated. With this information, we used KME to analyze the long-term survival of the DLPT scores for the cohort by stratifying data into subsets. R package `survminer` was used to produce the charts in this section.

1. Differences among Services

Figure 8 and Figure 9 show survival curves for listening and reading scores broken down by service component.

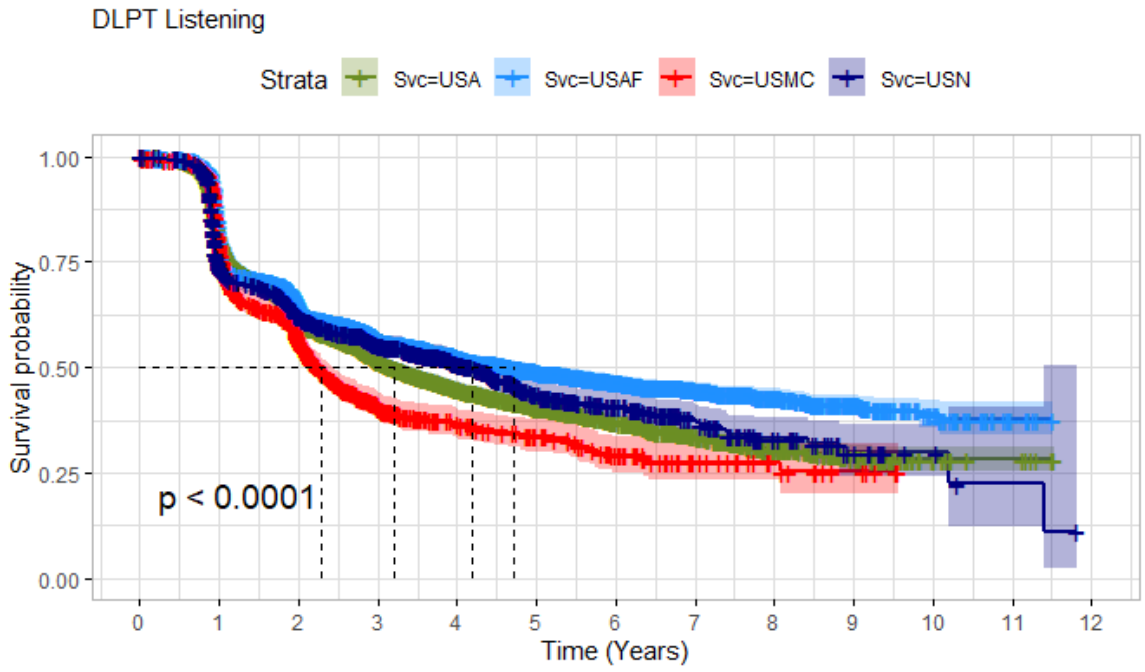


Figure 8. Survival Curves—by Service Component, DLPT Listening

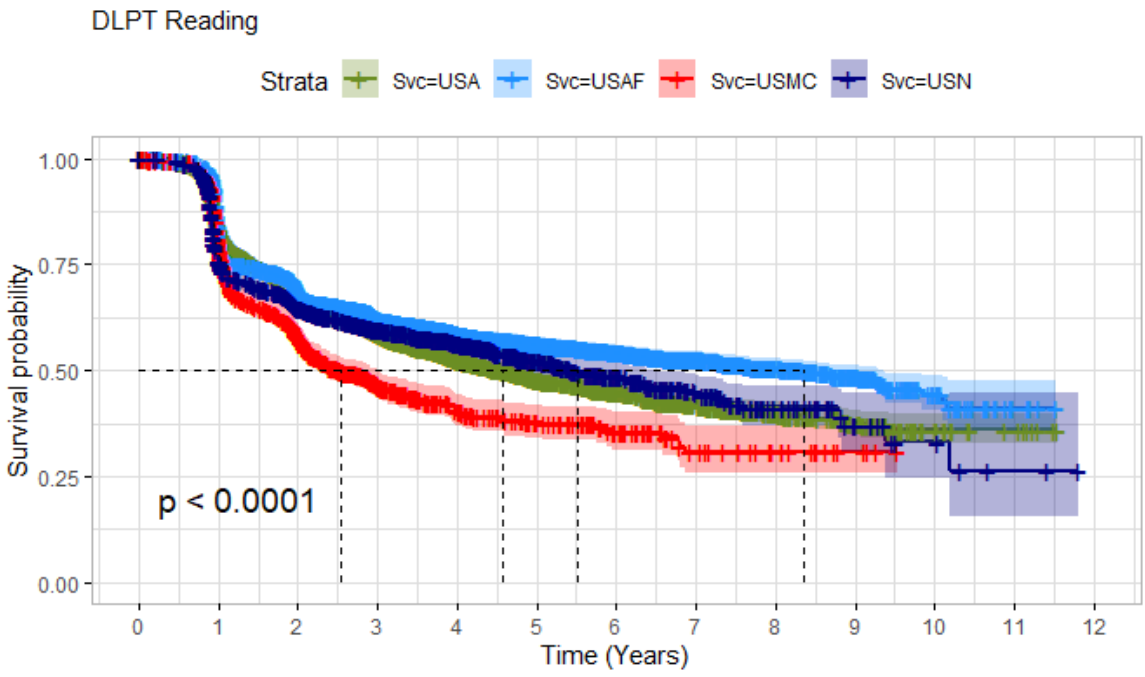


Figure 9. Survival Curves—by Service Component, DLPT Reading

Survival curves are statistically different from each other, in both listening and reading models for our dataset.

Table 10 shows the median survival times broken down by service component. Air Force students lead all other services in maintaining their DLPT scores achieved at graduation for both reading and listening test portions. This is in contrast to the one-year model, where Army scores maintain their scores at a higher rate. USMC students show the worst performance as a cohort.

Table 10. Median Survival Times by Service Component (years)

Service	Listening	Reading
USA	3.20	4.57
USAF	4.71	8.35
USMC	2.28	2.53
USN	4.19	5.52

2. Differences by GPA

In our one-year model, GPA is the most important predictor of maintaining a DLPT score. To show the difference the GPA makes in long-term score survival we binned this continuous variable into three categories: GPA below 3.4, GPA between 3.4 and 3.8, and GPA greater than 3.8. The results can be seen in Figure 10 and Figure 11.

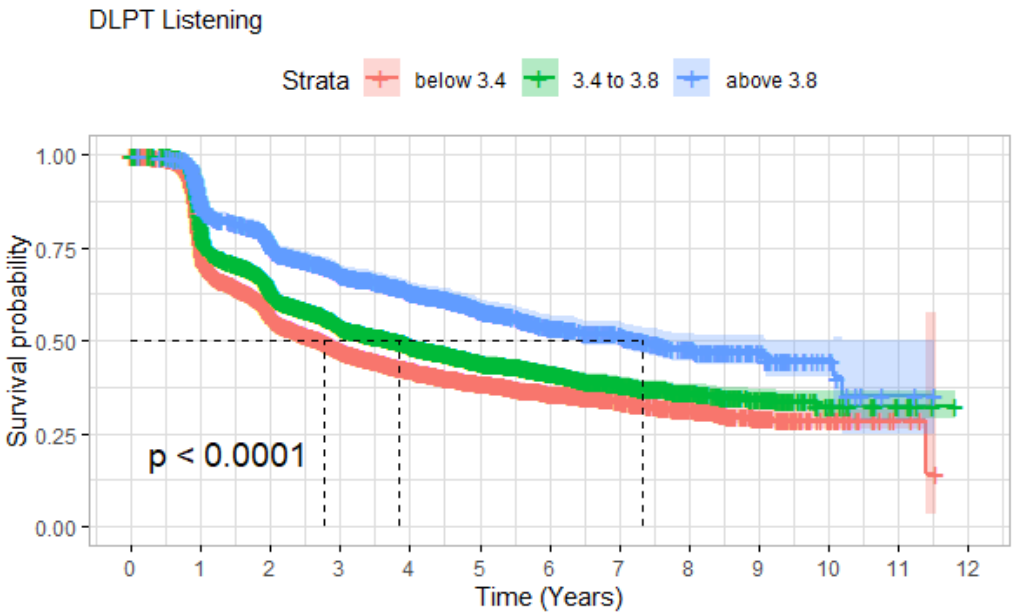


Figure 10. Survival Curves Separated by GPA, DLPT Listening

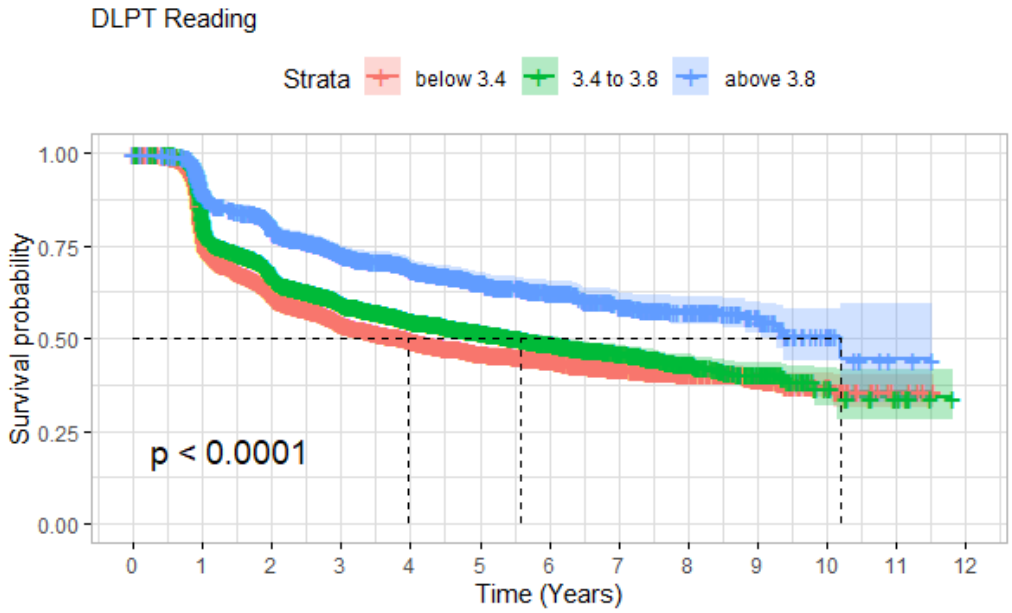


Figure 11. Survival Curves Separated by GPA, DLPT Reading.

Estimator shows statistical differences between groups, with higher GPA groups having better rates of survival (Table 11).

Table 11. Median Survival Times by GPA (years)

GPA	Listening	Reading
below 3.4	2.76	3.96
3.4 to 3.8	3.83	5.59
above 3.8	7.34	10.19

Consistent with our one-year model, the overall GPA is a discriminator in predicting long-term language skill survival. A higher GPA is a good predictor of students maintaining their DLPT scores longer.

3. Differences by DLPT

We used categorical DLPT score variables, separate for listening and reading, to show how survival rates differ for students in our cohort. Results can be observed in Figure 12 and Figure 13 and Table 12.

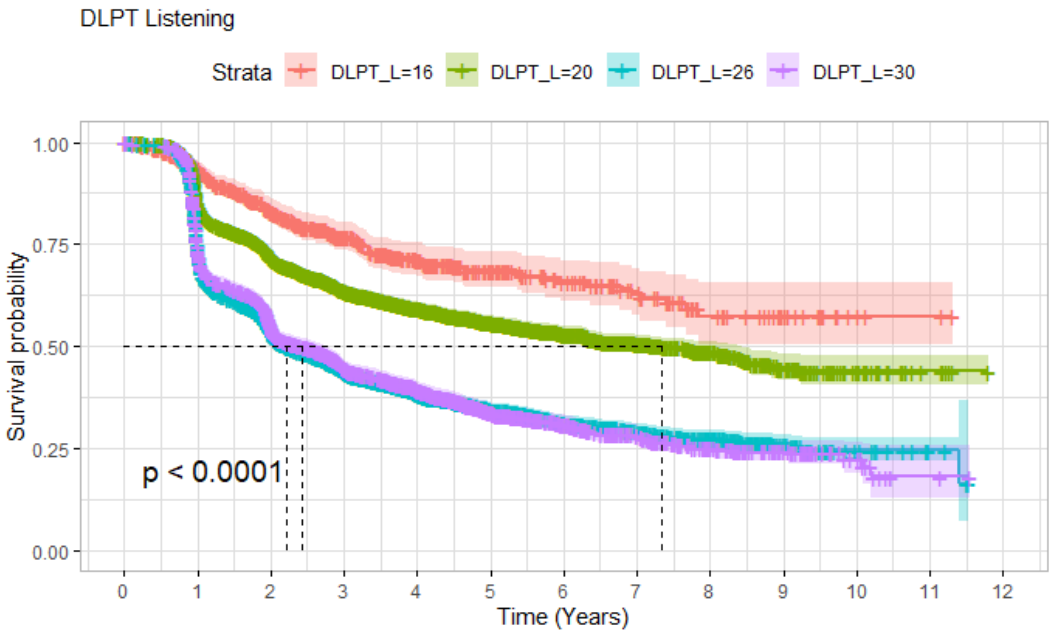


Figure 12. Survival Curves Separated by DLPT Scores, DLPT Listening

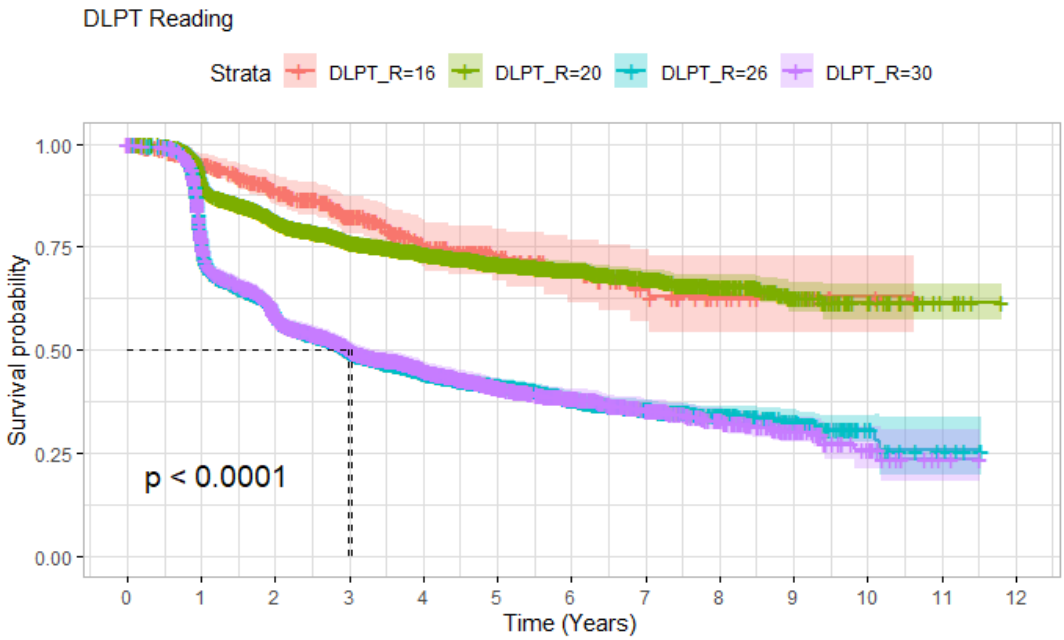


Figure 13. Survival Curves Separated by DLPT, DLPT Reading

Table 12. Median Survival Times by DLPT (years)

DLPT	Listening	Reading
1+	NA	NA
2	7.34	NA
2+	2.22	2.99
3	2.43	3.03

The long-term survival model supports the one-year model's findings—students that score higher on DLPT at graduation seem to score lower sooner than those who have lower scores at graduation.

4. Differences by Language Categories

Figures 14 and 15 show survival curves for listening and reading scores separated by language category. Category 1 languages are the easiest and Category 4 are the hardest to learn (Table 13).

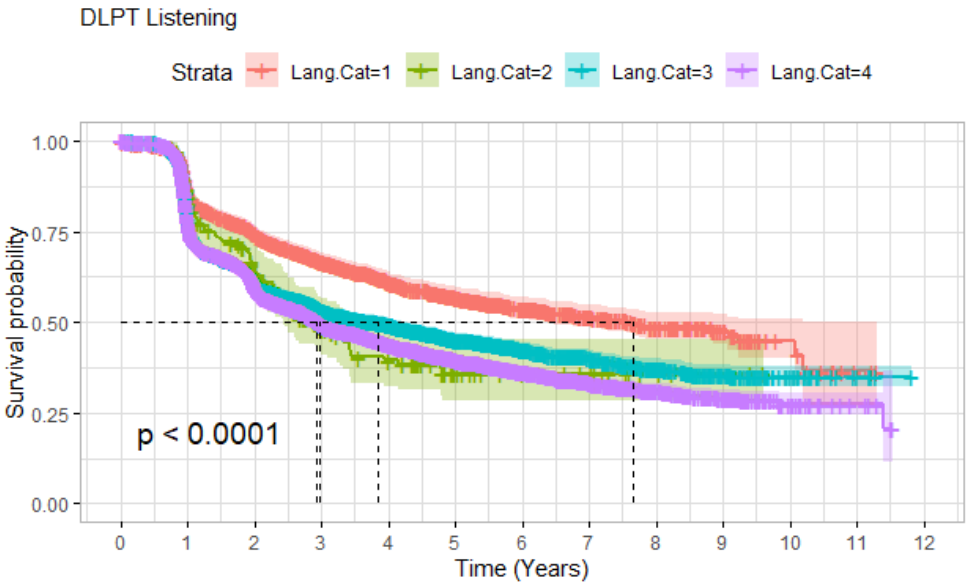


Figure 14. Survival Curves Separated by Language Category, DLPT Listening

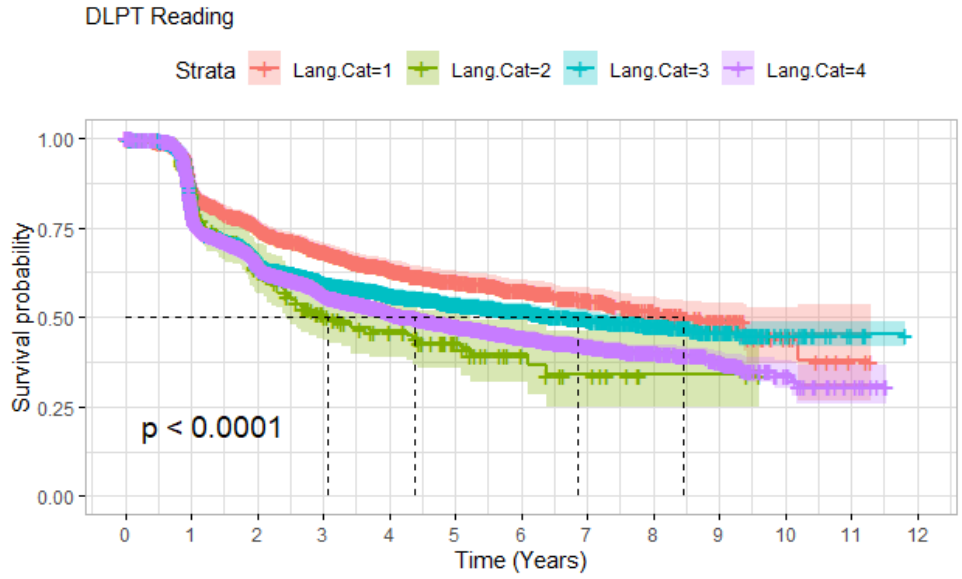


Figure 15. Survival Curves Separated by Language Category, DLPT Reading

Table 13. Median Survival Times by Language Category (years)

CAT	Listening	Reading
1	7.65	8.45
2	2.93	3.06
3	3.84	6.85
4	2.97	4.39

Category 1 languages show the best performance in terms of long-term survival rates, for both listening and reading scores. Category 3 languages perform better than Category 4, and Category 2 languages show the worst performance. This is in contrast to the one-year model, where differences between Category 1 and 2 languages were insignificant.

5. Immersion

The initial look at the Immersion Program showed that it improved the ability of students to maintain their scores longer. Figure 16 and Figure 17 show the differences in the survival curves for students who went through the Immersion program (of any kind), and students who had not had an opprtunity to participate (Table 14).

When we control for GPA, the picture changes, and shows no statistical differences for students who attended or did not attend the Immersion Program.

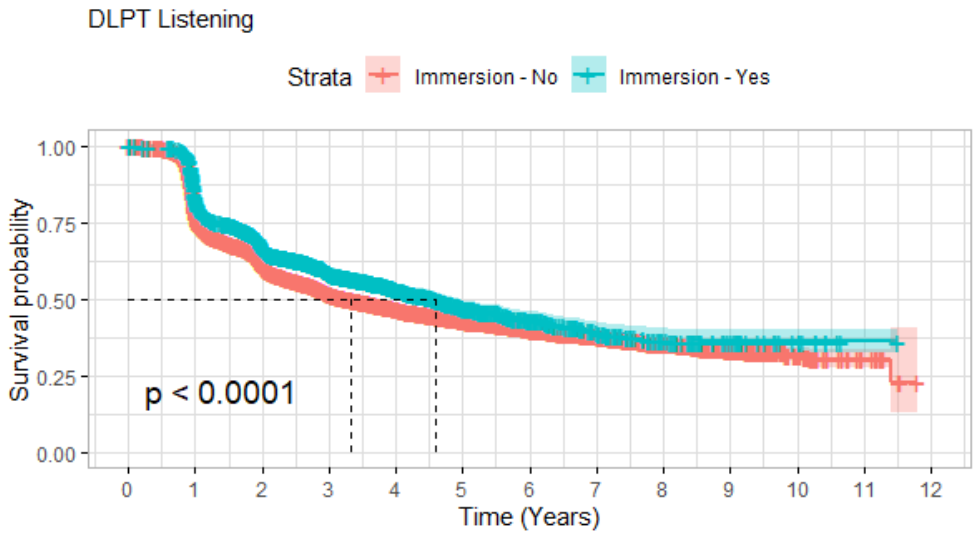


Figure 16. Survival Curves Separated by Immersion, DLPT Listening

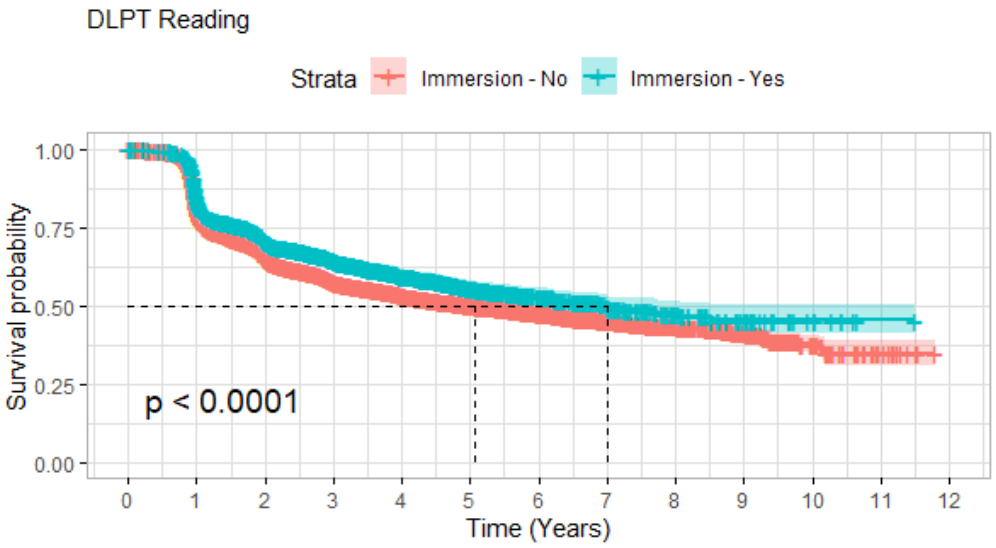


Figure 17. Survival Curves Separated by Immersion, DLPT Reading

Table 14. Median Survival Times by Immersion (years)

Immersion	Listening	Reading
Yes	4.59	7.00
No	3.33	5.07

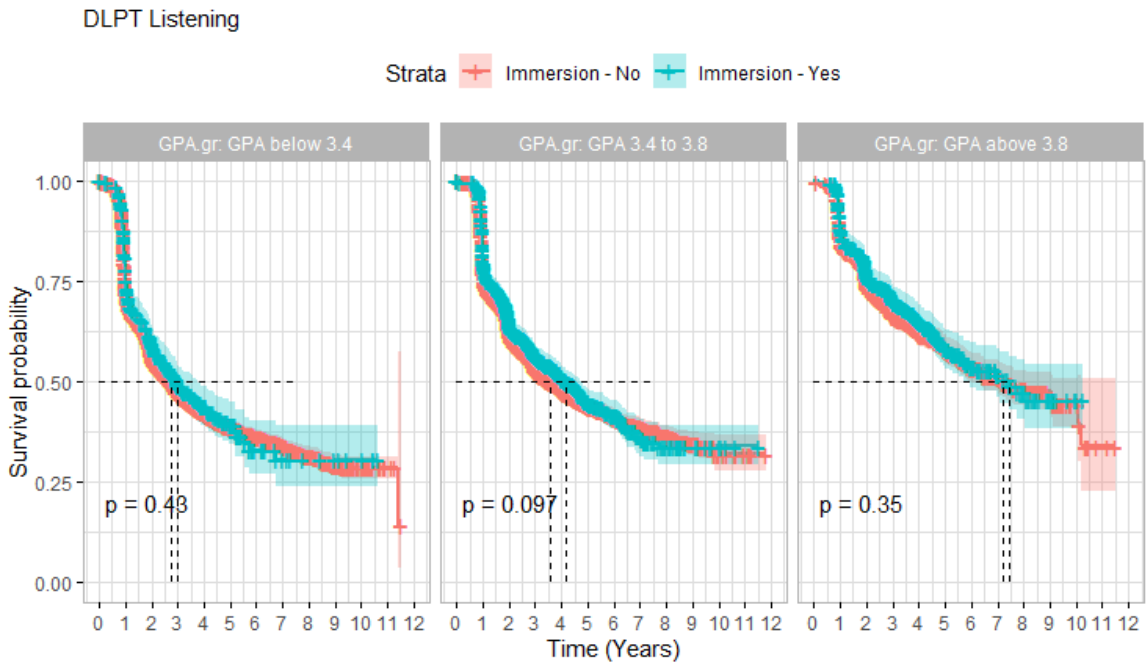


Figure 18. Survival Curves Separated by Immersion, Controlled for GPA, DLPT Listening

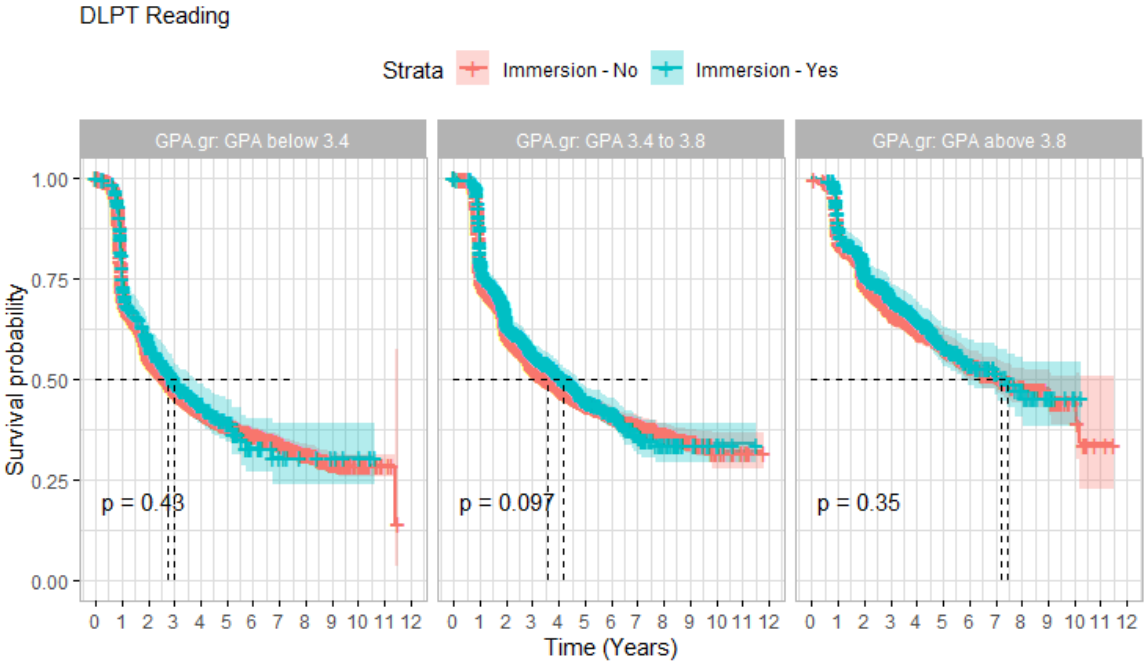


Figure 19. Survival Curves Separated by Immersion, Controlled for GPA, DLPT Reading

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSIONS

A. SUMMARY

The purpose of this study was to identify factors associated with the atrophy of acquired language skills for successful DLIFLC graduates over time. The objective was to gain a better understanding and potentially mitigate the atrophy drivers. We used a mixed approach that involved two different techniques to conduct our research. After shaping the problem as a survival analysis, we were able to observe that a significant drop in both reading and listening DLPT scores happens within the first year after graduation, likely during the first annual DLPT retest. In fact, the data showed that only 75.5% of the listening scores and 78.2% of reading scores survived past the first year unchanged.

The logistic regression model was fitted onto data to predict whether the score would drop within the first year. We found that several factors were associated with maintaining the scores after graduation, with the overall GPA being the most important predictor of success. The surprise finding was that students who scored higher on the DLPT at graduation had higher odds of dropping the scores during the first postgraduation retest. The Army students, students who studied easier languages, senior non-commissioned officers and officers, and students who went through the OCONUS immersion program had better odds of maintaining their scores. A number of variables were not significant predictors and were eliminated from the model.

We conducted an additional, long-term survival analysis using the understanding of the factors from the one-year model. The long-term survival probabilities in general supported the one-year model findings, with some exceptions. The Army students have shown the best survival rates for both scores in the short-term model, but the Air Force students had better survival probabilities in the long run.

B. FUTURE WORK

This study was somewhat limited in its scope because it did not take into account the events that happen after a student graduates from DLIFLC. The types of assignments and specific jobs that students perform in the operating force have various levels of need

for language use in day-to-day duties, and likely have some influence on how language skills change over time. Administrative problems, changes in marital status, and additional dependents can potentially affect how language skills atrophy over time. Another possible factor is the additional language training that service members receive over time after graduating from DLIFLC, including training in non-DOD schools. DMDC collects all these data, and DLIFLC can obtain it to incorporate into a future study. Expanding the current study by including post-DLIFLC events would create a better understanding of the learned language longevity and ways to improve it.

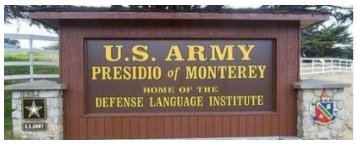
Another area of interest is the retention of the service members that have undergone language training at DLIFLC. All the factors listed above, combined with data that was used in the study can be exploited to determine what factors influence junior enlisted members' decisions to stay in or leave the service after the first enlistment.

APPENDIX A. DLIFLC ESTIMATED TRAINING COST PER GRADUATE (FY 2019)

Resident Training Cost per Graduate

FY18 Training / FY19 Dollars

Presidio of Monterey, CA	**64.0**	**Course Length (weeks)**
Defense Language Institute		**Flying Hours**
ARABIC BASIC PEP	**E-4**	**Modal Grade**
01AD-P	**98**	**Equivalent Graduates**
	22.0	**Frequency**
	3,840	**ICH**



	Mil Pay	Civ	OMA NP	Other	Total
DIRECT COSTS					
1. Instruction	10,720	128,254	419		139,392
2. Flying Costs					
3. Overhead	8,708	21,693	31,527		61,929
4. Ammunition					
5. Sub-Total	19,428	149,947	31,946		201,321
6. Student Costs					
Pay & Allowances	86,229				86,229
Per Diem					
Travel	10,318				10,318
7. Sub-Total	96,548				96,548
8. TOTAL DIRECT	115,976	149,947	31,946		297,869
INDIRECT COSTS					
9. Base Support	648	2,209	7,956	791	11,604
10. Medical Support	3,180	6,418	3,465		13,063
11. Family Housing					
12. TOTAL INDIRECT	3,828	8,627	11,421	791	24,668
13. TOTAL DIR. + INDIR.	119,804	158,574	43,367	791	322,536

Fixed & Variable Costs					
14. DIRECT					
a. Fixed	7,847	57,226	14,090		79,163
b. Variable	108,128	92,721	17,856		218,705
15. DIRECT + INDIRECT					
a. Fixed	10,067	62,562	20,803	465	93,896
b. Variable	109,737	96,012	22,565	326	228,640
			TOTAL EGRAD	**per**	**$322,536**

NOT FOR USE IN PROGRAMMING/BUDGETING DRILLS

SOURCE: ATRM-PDC (DSN 501-6705)

APPENDIX B. DLIFLC LANGUAGES IN THE DATASET

Language Name	Three-Letter Code	Two-Letter Code	Category	Taught at DLIFLC FY2022
ARABIC-MODERN STANDARD	ARB	AD	4	Yes
ARABIC-EGYPTIAN	ARZ	AE	4	Yes
ARABIC-GULF	QAG	DG	4	Yes
ARABIC-SAUDI	QAS	AP	4	Yes
ARABIC-SUDANESE	APD	AV	4	No
CHINESE-MANDARIN	CMN	CM	4	Yes
FRENCH	FRA	FR	1	Yes
GERMAN	DEU	GM	2	No
HEBREW	HEB	HE	3	Yes
HINDI	HIN	HJ	3	No
INDONESIAN	IND	JN	2	Yes
ITALIAN	ITA	JT	1	No
JAPANESE	JPN	JA	4	Yes
KOREAN	KOR	KP	4	Yes
PERSIAN-AFGHAN	PRS	PG	3	No
PERSIAN-IRANIAN	PES	PF	3	Yes
PORTUGUESE	POR	PY	1	No
PUNJABI	PAN	PJ	3	No
PUSHTU-AFGHAN	PBT	PV	4	Yes
RUSSIAN	RUS	RU	3	Yes
SERBO-CROATIAN	HBS	SC	3	No
SPANISH	SPA	QB	1	Yes
TAGALOG	TGL	TA	3	Yes
THAI	THA	TH	3	No
TURKISH	TUR	TU	3	No
URDU	URD	UR	3	Yes
UZBEK	UZB	UX	3	No

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Aalen BO, Gjessing H. (2008) *Survival and Event History Analysis: A Process Point of View*. (Springer, New York, NY).

Bermudez-Mendez J (2020) Student success factors at Defense Language Institute Foreign Language Center. Master's thesis, Operations Research Department, Naval Postgraduate School, Monterey, CA, https://calhoun.nps.edu/handle/10945/64866

Brenner IA (2021) Student achievement indicators at the Defense Language Institute Foreign Language Center. Master's thesis, Operations Research Department, Naval Postgraduate School, Monterey, CA

Defense Language Institute Foreign Language Center (2017) Institutional self-evaluation Report, DLIFLC, Monterey, CA. https://www.dliflc.edu/wp-content/uploads/2018/01/DLIFLC-Self-Study-December-2017_small.pdf.

Defense Language Institute Foreign Language Center (2019) General Catalog 2019–2020, v10c (Monterey, CA). https://www.dliflc.edu/wp-content/uploads/2018/11/DLIFLC_catalog_2019-20_v10c.pdf.

Department of the Army (2015) Defense Language Institute Foreign Language Center (DLIFLC) plan to achieve 2+/2+ executive summary. Memorandum, Washington, DC.

Department of Defense (2009) DOD language testing program. DOD Instruction 5160.71, Washington, DC, https://dlnseo.org/sites/default/files/DoDI_5160.71.pdf.

Géron A (2019) *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* (O'Reilly Media, Inc., Sebastopol, California).

James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning with Applications in R.* (Springer, New York)

Haupt AC (2014) Analysis of Korean academic attrition at the Defense Language Institute Foreign Language Center. Master's thesis, Operations Research Department, Naval Postgraduate School, Monterey, California.

Hinson WB (2005) A statistical analysis of individual success after successful completion of Defense Language Institute Foreign Language Center Training. Master's thesis, Naval Postgraduate School, Monterey, California.

Hosmer DW, Lemeshow S, Sturdivant RX (2013) *Applied Logistic Regression* Third Edition (John Wiley & Sons, Hoboken, NJ).

Kim, JW, Ritter, FE, Koubek, RJ (2011) An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science* 1–16. doi:10.1080/1464536X.2011.573008

R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, http://www.R-project.org

Rich, JT, Neely, JG, Paniello, RC, Voelker, CC, Nussenbaum, B, Wang, EW (2010) A practical guide to understanding Kaplan-Meier curves. Otolaryngology—Head and Neck Surgery: *Official Journal of American Academy of Otolaryngology-Head and Neck Surgery* 143(3), 331–336. https://doi.org/10.1016/j.otohns.2010.05.007

Shearer SR (2013) Modeling second language change using skill retention theory. Doctoral dissertation, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/34742

Wong CH (2004) An Analysis of Factors Predicting Graduation of Students at Defense Language Institute Foreign Language Center. Master's thesis, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/1296

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California