
AFRL-RH-WP-TR-2021-0012

# Air Force Reading Abilities Test: Revisions and Additions

**Daniela A. Pena**
**Luisa Martinez**
**Natasha Haight**
**C. Wayne Shore**

**Operational Technologies Corporation**

**February 2021**
**Interim Report**



**AIR FORCE RESEARCH LABORATORY**
**711TH HUMAN PERFORMANCE WING,**
**AIRMAN SYSTEMS DIRECTORATE,**
**WRIGHT-PATTERSON AIR FORCE BASE, OH 45433**
**AIR FORCE MATERIEL COMMAND**
**UNITED STATES AIR FORCE**

**NOTICE AND SIGNATURE PAGE**

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2021-0012 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

CARRETTA.THOMAS.R.1228984929
Digitally signed by CARRETTA.THOMAS.R.1228984929
Date: 2021.03.31 09:41:34 -04'00'

THOMAS R. CARRETTA
Work Unit Manager
Performance Optimization Branch
Airman Biosciences Division

MCKINLEY.RICHARD.A.1262239710
Digitally signed by MCKINLEY.RICHARD.A.1262239710
Date: 2021.03.31 13:23:28 -04'00'

R. ANDY MCKINLEY, DR-III, PhD
Core Research Area Lead
Performance Optimization Branch
Airman Biosciences Division

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE *(DD-MM-YY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| 26/02/21 | Interim | 08/20/2018 -02/26/2021 |

**4. TITLE AND SUBTITLE**

Air Force Reading Abilities Test: Revisions and Additions

**5a. CONTRACT NUMBER**
F8650-14-D-6500/FA8650-18-F-6828

**5b. GRANT NUMBER**
Not applicable

**5c. PROGRAM ELEMENT NUMBER**
62202F

**6. AUTHOR(S)**
Daniela A. Pena, Luisa Martinez, Natasha Haight, & C. Wayne Shore

**5d. PROJECT NUMBER**
5329

**5e. TASK NUMBER**
09

**5f. WORK UNIT NUMBER**
H0SA (532909TC)

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Operational Technologies Corporation
4100 N.W Loop 410
Suite 100
San Antonio, Texas 78229

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Airman Biosciences Division
Performance Optimization Branch
Wright-Patterson AFB, OH 45433

Infoscitex Corporation
4027 Colonel Glenn Highway
Suite 210
Dayton, OH 45431-1672

**10. SPONSORING/MONITORING AGENCY ACRONYM(S)**
711 HPW/RHBC

**11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)**
AFRL-RH-WP-TR-2021-0012

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
Distribution Statement A: Approved for public release. AFRL-2021-1027, cleared 30 March 2021

**13. SUPPLEMENTARY NOTES**
Work performed under subcontract number FPH02-S027, PO 182140. . Report contains color.

**14. ABSTRACT**

The purpose of this study was to update the Air Force Reading Abilities Test (AFRAT) with content that meets current standards, develop two parallel forms, and equate them to a previous version and to its Reading Grade Level scale. A second goal of this study was to develop an Orthographic and a Phonological Choice test. The combined use of these tests is intended to identify Airmen with potential reading disabilities so that they can receive further assessment or be referred to remediation programs. The tests were administered to Air Force Basic Military Trainees (BMTs) and to Amazon Mechanical Turk (MTurk) workers. All participants received one of the two new reading comprehension forms and all of the other instruments. The new tests were found to be parallel to each other and the previous form, and composite scores were equated to the previous form using equipercentile equating. The new AFRAT subtests show appropriate validity and reliability and appear to be an effective tool for flagging potential reading disabilities.

**15. SUBJECT TERMS**
Air Force Reading Abilities Test, reading grade level, reading disorders

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT: | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON (Monitor) |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Thomas R. Carretta |
| Unclassified | Unclassified | Unclassified | SAR | 43 | 19b. TELEPHONE NUMBER *(Include Area Code)* |

**TABLE OF CONTENTS**

## LIST OF FIGURES



## LIST OF TABLES

**ACKNOWLEDGEMENTS**

We would like to sincerely thank the following individuals for their valuable contributions to this effort.

Thomas R. Carretta, PhD (Contributor)
Miguel Gonzalez (Contributor)
Malcolm J. Ree, PhD (Consultant)
Amanda Mouton (Editor)
Andrew Deregla (Editor)
John Trent (Contributor)
Benjamin Fairbank, PhD (Editor)
Kristina Breaux, PhD (Subject Matter Expert)

**SUMMARY**

The purpose of this study was to develop two new versions of the reading comprehension subtests of the Air Force Reading Abilities Test (AFRAT), plus two additional sections meant to identify Airmen with potential reading disabilities. The current AFRAT is used to identify service members who have difficulty with reading and to direct them to remedial training programs. The AFRAT provides a standardized Air Force-wide Reading Grade Level (RGL) measure.

The new forms were equated to the previous AFRAT form (Form 81A) and its accompanying RGL scale through equipercentile equating. We used a counterbalanced single group design. Equated scores were then converted to a RGL scale for the total scores. Data were collected using participants from two population sources: Air Force Basic Military Trainees (BMTs) and Amazon's Mechanical Turk (MTurk) workers.

The Orthographic and Phonological Choice subtests are speeded tests that were developed as a new section to the AFRAT. They were designed to identify Airmen with a potential reading disability such as dyslexia.

Examination of item statistics showed that the new Reading Comprehension forms are parallel to each other and to the previous versions. The new versions had an average item difficulty of .80 and an average biserial correlation of .56. The previous versions had an average difficulty of .84 and a biserial correlation of .63.

Analyses of the Orthographic and Phonological Choice subtests indicated that low scores on these tests are related to lower RGL and to increased risk of having a reading disability as measured by the Adult Reading History Questionnaire (ARHQ). Low percentile scores on these subtests identify Airmen most likely to have some form of reading disability. Data collection from a larger Air Force sample, along with the accumulated knowledge of the identification of reading disabilities, will allow for the determination of a standard score that accurately identifies reading deficiencies. We have presented ways in which these tests, along with the RGL measure, can identify Airmen who should receive additional assessment and/or training.

The ARHQ, a screening tool for dyslexia, was used as the best available validation of the new versions of the AFRAT and the new subtests. The correlation between the ARHQ and the new forms provided adequate evidence of validity.

We recommend that the two revised versions of the reading comprehension forms and the two newly created subtests for specific reading issues be used together to identify Airmen that potentially have a reading disability that may impair their ability to function as effective Airmen, whether in training or on the job.

# 1.0 INTRODUCTION

Many Air Force Specialties (AFSs) include duties that require a specified level of competency in reading. Technical training often involves extensive reading even when the duties of the AFSs themselves do not emphasize it. Low levels of reading comprehension can impact training success, with dyslexia being a particular concern. Eighty percent of individuals with learning disabilities have been identified as having dyslexia, making it the most common learning disability (Shaywitz, Gruen, & Shaywitz, 2007). College students with dyslexia have reported struggles with note taking, organizing essays, writing, remembering facts, listening, concentrating, time keeping, and other learning skills (Mortimore & Crozier, 2006).

Reading deficiencies can be addressed and often managed with interventions which have been shown to improve educational success (Wanzek, Wexler, Vaughn, & Ciullo, 2010; Woodruff, Schumaker, & Deshler, 2002). A specialized reading ability assessment is necessary to determine whether training issues are caused by reading deficiencies. The Air Force Reading Abilities Test (AFRAT) was designed to make such determinations. The current research effort updated the reading comprehension section of the AFRAT and developed two new tests to better identify members with potential reading disabilities.

## 1.1 Background

### 1.1.1. AFRAT

The AFRAT was produced by the Air Force Human Resources Laboratory (AFHRL) to replace an extensive list of commercial tests that were previously used by the Air Force. The developers of the AFRAT sought to produce a reading comprehension test with norms appropriate for military populations and to standardize reading ability judgments across the Air Force. Reading Grade Level (RGL) varied widely for individuals with similar Armed Services Vocational Aptitude Battery (ASVAB) scores depending on the reading measurement administered (Mathews, Valentine, & Sellman, 1978). The AFRAT would cease the use of commercialized reading tests (Riemer, 1984). Two forms were designed (Form 81A and 81B), with separate sections for vocabulary and reading comprehension (Mathews & Roach, 1983). The capability of the AFRAT was first assessed using a test form composed of preexisting items, which became Form 82X. Forms 81A and 81B were produced by drawing items from a large item pool written by the Educational Testing Service (Massey & Matthews, 1980). They were validated against commercial reading tests commonly used in the Air Force (Mathews & Roach, 1983). Form 82X was shown to have predictive validity for technical training grades (Mathews & Roach, 1983).

The AFRAT contains 45 vocabulary items and 40 reading comprehension items. The reading comprehension items were structured as a reading passage composed of one or more paragraphs followed by at least one item referring to the passage. The items were multiple choice with four alternatives each. Participants were required to paraphrase the passage's content or make inferences from it. The reading comprehension portion of the test was limited to 35 minutes out of the 50 minutes allowed for the entire test.

The AFRAT was implemented in 1982 and has been used Air Force wide as a RGL information tool. More specifically, it is commonly used to identify if Airmen who experience training difficulties do so because they read at a level lower than what is required to be successful in their career field. After completing the assessment, these trainees may be directed to mandatory or voluntary remedial training. Because the test was designed to identify reading deficiencies, most

of its evaluative power lies in identifying individuals with low reading abilities (Mathews & Roach, 1983). Conversion tables transform AFRAT scores to RGLs ranging from the fourth-grade level to the twelfth-grade level (Mathews & Roach, 1983). Additional information on the AFRAT can be found in Skinner, Thompson, Schwartz, and Weissmuller (2007).

### 1.1.2. Reading Disability

Dyslexia is marked by a particular difficulty in reading for both children and adults who otherwise possess the intelligence, motivation, and learning opportunities to be proficient readers (Shaywitz, 1998). These reading difficulties have been found to manifest primarily as difficulty identifying written words (Vellutino, Fletcher, Snowling, & Scanlon, 2004). While dyslexia was historically thought to be caused by a disability of visual perception and memory systems, research has shown that it is more likely rooted in the storage and retrieval of linguistic information (Vellutino, 1987). Likewise, the theory that deficits in eye-tracking of visual stimuli lead to dyslexia has been refuted, as well as have other theories attributing dyslexia to deficits in basic learning processes (Vellutino et al., 2004). Extensive research has supported poor phonological processing skills (i.e., managing the basic sounds that make up words in such ways as recognizing, decoding, encoding, combining, and separating them) as the underlying cause of dyslexia (Grigorenko, 2001; Vellutino et al., 2004). Individuals with weak phonological awareness have been found to also have inadequate orthographic knowledge (i.e., understanding the rules that govern how letters may be arranged in words; Vellutino et al., 2004). Even dyslexic individuals who apparently read at a normal level have demonstrated difficulty with spelling (Treiman, 1997).

While dyslexia is associated with difficulty recognizing written words (Vellutino et al., 2004), another disability, specific reading comprehension deficit (S-RCD), involves adequate or superior word recognition but difficulty comprehending what is read (American Speech-Language-Hearing Association; ASHA, n.d.; Landi & Ryherd, 2017). Deficits in writing or reading can have a circular relationship, and other language disorders can produce similar outcomes as reading disorders (ASHA, n.d.). Reading troubles can be caused by deficits in a variety of abilities, and thus any one-dimensional test is inadequate to fully determine the nature of low RGL for a given individual. Specialized testing would be needed to provide better information on the type of intervention that will be most beneficial.

## 1.2 Purpose

The current effort endeavored to enhance the effectiveness of the AFRAT for Air Force populations by updating the reading comprehension passages and items. The previous reading comprehension passages contained general information that is not relevant for the Air Force and outdated language that does not meet current standards. An objective of this project was to update the reading comprehension items by using paragraphs with content relevant to Air Force trainees. Additionally, a new section of the AFRAT was developed to identify Airmen with potential, unidentified reading disabilities. The Air Force believes that individuals who experience difficulty with reading will benefit from interventions that address their reading deficiencies.

# 2.0. APPROACH AND METHODOLOGY

## 2.1 Design Goals

The AFRAT was designed with the following goals (Riemer, 1984):

1. Contain vocabulary and reading comprehension sections, as in the Test of Adult Basic Education (TABE) and similar tests.
2. Compose reading comprehension items from expository prose.
3. Compose reading comprehension items from text that is factual, but unlikely to be answered correctly from previous knowledge.
4. Construct vocabulary items from words likely to be encountered in the Air Force.
5. Maximize reliability while maintaining a time limit of under one hour.

The vocabulary items were judged to be current so the goals relevant to the current project are numbers 2, 3, and 5 above. These goals served as reference points when developing the updated reading comprehension items for the test. Items were designed to include expository and factual reading content, to avoid commonly known information, and to limit test administration time. At the request of the Air Force, all passages used in the updated AFRAT came from Air Force training materials to ensure a close match between the content of the AFRAT and reading content actually encountered by Air Force service members in training.

## 2.2 Test Development

### 2.2.1. Development of New AFRAT Reading Comprehension Forms

To produce updated reading comprehension items that retain the essential characteristics of the original items, we reviewed the AFRAT's descriptive statistics, including means, standard deviations, and item difficulties. These statistics were matched as closely as possible in the new test forms without diminishing their discriminatory capabilities.

Content for reading comprehension items was chosen from Non-Commissioned Officer (NCO) Academy materials furnished by the Air Force Personnel Center (AFPC)/Strategic Research and Assessment Branch (DSYX). The passages that were the most information rich and of moderate length (129 to 406 words) were chosen, as these were deemed most conducive for producing an adequate number of items (3 to 8 per passage) without dominating test time. Passages were altered slightly to enhance their readability as stand-alone excerpts. Additionally, if the RGL of the passage was judged to be too high for the purposes of the AFRAT, adjustments were made to lower the difficulty of the passage (i.e., use more common synonyms for words, simplify sentence structure, and shorten sentence length). Items were generated from these passages using a taxonomy of six types of reading comprehension questions (see Table 1) commonly used on examinations such as the Scholastic Aptitude Test, Graduate Record Exam, Law School Admissions Test, and Air Force Officer Qualifying Test. Since a taxonomy was not found for the previous AFRAT, its items were reviewed and assigned to one of these six taxonomic categories. This was done to ensure greater distribution of items developed across taxonomic categories compared to the previous AFRAT. Table 1 shows the number of items from each taxonomic category for the three AFRAT reading comprehension forms.

**Table 1. AFRAT Taxonomy**

| Taxonomy | 81A (Old) | Form A (New) | Form B (New) |
|---|---|---|---|
| Summary | 6 | 8 | 8 |
| Stated (reworded) details | 29 | 13 | 11 |
| Underlying assumptions | 2 | 7 | 8 |
| Rhetorical function(s) of portions of the text | 0 | 6 | 8 |
| Author's viewpoint/ Inferred statements | 2 | 5 | 2 |
| Analogy | 1 | 1 | 3 |

Newly written items were first reviewed for quality, fairness, and bias. AFPC/DSYX personnel provided additional review. Accepted items were compiled into test forms and administered to enlisted personnel attending Basic Military Training (BMT) at the Applied Performance and Testing (APAT) center at Lackland Air Force Base, Texas. After calculating item and difficulty statistics, items found to have misleading distractors or other issues were either removed or revised and re-tested. Items with biserial correlations between .32 and .80 and difficulties between .51 and .92 were accepted for use in the final test forms. These ranges were chosen to approximate the original AFRAT forms. The average difficulty level in the original AFRAT was .83 for both form 81A and 81B, and the average biserial correlations were .61 for form 81A and .65 for form 81B (Mathews & Roach, 1983). The ranges of the difficulty values for the original items were from a lower limit of .59 or lower to an upper limit of .90 or above, with the majority of the items being above .80. The biserial correlations of the original AFRAT forms ranged from .30 to .89 (Mathews & Roach, 1983). The newly developed items had a slightly higher difficulty range and had a slightly lower maximum biserial correlations. This was done to ensure that the final test forms had the desired average difficulty, to balance items across the taxonomic categories, and to prevent some items from being so easy that they could be answered correctly by almost all examinees regardless of their RGL. To make the new versions parallel to one another, careful consideration went into balancing word length of the passages. Form A had a total word count of 1,669 words (3,128 counting the items) and Form B had a total of 1,565 words (3,142 counting the items). Item difficulty, biserial correlation, and taxonomic category were also considered when balancing the final versions of the tests.

### 2.2.2. Orthographic and Phonological Choice Test Development

To produce a reading disability screening test that can be used broadly and cost-effectively in the Air Force, a scrambled letters prototype test was suggested by AFPC/DSYX as a new addition to the AFRAT. After consultation with a subject matter expert (SME) in clinical assessment of Dyslexia, Dr. Kristina Breaux, Senior Research Director at Pearson Education, it was concluded that the scrambled letters test would not be appropriate. Research has not supported the ability of the test to discriminate between individuals who may have a reading disability and those who do not. Additionally, the SME stated that the Scrambled Letters test involves paired associate learning with a visual response. People with dyslexia are not impaired in this area, so the test would have been contaminated by an irrelevant construct. Following the advice of the SME, the Orthographic and Phonological Choice tests were selected and developed. The SME noted that

if only one test could be developed, the Phonological Choice test was the better option. We had enough resources to develop both.

In addition to being supported by research, the Orthographic and Phonological Choice tests are easy to score and administer to many examinees simultaneously (Wolff & Lundberg, 2003). In the Phonological Choice test, examinees are shown three nonsense words (e.g., baybee, kiddle, and geegum), and are instructed to choose the one that is pronounced like a real word (e.g., baybee). In the Orthographic Choice, examinees are shown three alternate spellings for a word (e.g., scelaton, skeleton, and skelletan) and are instructed to select the one that is spelled correctly (e.g., skeleton). Both the Phonological and Orthographic Choice tests discriminated well between adults with and without dyslexia, outperforming several other indicators in a battery of tests (Wolff & Lundberg, 2003). Bruck (1992) found that adults with dyslexia never reached age-appropriate levels of phoneme awareness (i.e., awareness of the individual sounds that make up spoken words). Orthographic awareness (i.e., awareness of the rules by which letters are arranged in words) was also found to be weaker in adult dyslexics than a control group (Kemp, Parrila, & Kirby, 2008). No gender differences were observed in the Phonological Choice test for individuals with or without dyslexia, but dyslexic females performed significantly better than dyslexic males on the Orthographic Choice test (Wolff & Lundberg, 2003).

The SME advised that the test must be administered under a time constraint to adequately discriminate between individuals with and without potential reading disabilities. In addition, the SME recommended penalizing guessing by subtracting the number of incorrect answers from the number of correct answers. The SME cautioned that these tests alone were not sufficient to identify dyslexia, but rather could serve as an indicator or flag for individuals who might be at risk. Additional testing by health care professionals would be necessary for confirmation. For the purposes of the AFRAT, these parameters are acceptable. The AFRAT was not designed to diagnose reading disability disorders, but to provide an indicator that additional training or consultations are needed. A report by the National Center on Adult Literacy concluded that in practice, reading training designed for adults with dyslexia was also effective for non-dyslexic adults struggling with reading (Fowler & Scarborough, 1993). The report emphasized that instruction should focus on the specific underlying skills that were deficient in an individual. Based on this conclusion, a measure that indicates symptoms consistent with reading disabilities does not risk misclassifying non-disabled individuals into training that would not benefit them; both those with and without a reading disability benefit similarly. Orthographic and Phonological items were developed according to the SME's guidelines and were then tested at the APAT facility. Item difficulty and discrimination statistics were calculated for both. Different time limits were tested (two minutes vs three minutes) to ensure it was a speeded test. The resulting data indicated that giving examinees two minutes to complete each test was optimal.

As suggested by the SME, we used a self-report screening tool for risk of reading disabilities to establish construct validity of the Orthographic and Phonological Choice tests. The Adult Reading History Questionnaire (ARHQ; Lefly & Pennington, 2000) was added for this purpose. The ARHQ is a self-report instrument in which participants rate their experience and preferences with reading. The ARHQ has been found to have good validity and to effectively discriminate between adults with and without reading disabilities (Lelfly & Pennington, 2000; Welcome & Meza, 2019). We removed two items from the ARHQ due to redundancy of content. The modified ARHQ included 20 items in a Likert scale response format ranging from 0-4, and three Yes/No questions asking participants about their reading history (see Appendix A). The

ARHQ was scored by adding the total points for each participant and dividing them by the total number of possible points, with higher scores indicating more reading difficulty risk.

## 2.3 Sampling Procedure

Participants completed five instruments: one of the new forms of the AFRAT, AFRAT Form 81A (with both the reading comprehension and vocabulary subtests), the Orthographic and Phonological Choice test, the ARHQ, and demographic questions. Total testing time was about 100 minutes.

To counteract fatigue, boredom, or other order effects, a counterbalanced approach was used. The order in which participants viewed the old and new reading comprehension tests was split into four test conditions. In two conditions, participants received one of the new versions of the test first (i.e., form A or form B), the vocabulary subtest, followed by the old version (i.e., form 81A), the Orthographic and Phonological Choice test, and lastly the ARHQ. In the other two conditions, participants received the old version first, and one of the new versions last, with the order of the Vocabulary, Orthographic and Phonological Choice tests, and the ARHQ unchanged.

### 2.3.1. APAT Procedure

Data from an Air Force sample were collected through the APAT center. At this center, experimental Air Force tests are given to BMTs in person. The new forms of the AFRAT and its accompanying instruments were administered to trainees in a paper-and-pencil format by trained proctors and according to standardized instructions. Participants recorded their answers on answer sheets for all sections except the Orthographic and Phonological Choice subtests. Due to the strict time limits imposed for these two subtests, answers were provided directly on the test booklet. This was also done to minimize differences in performance between the paper-and-pencil format and the computerized version that was administered using Amazon's Mechanical Turk (MTurk). Participants were randomly assigned into one of the four counterbalancing conditions previously mentioned.

### 2.3.2. MTurk Procedure

Data from a civilian sample were collected using MTurk. MTurk is an internet-based crowdsourcing platform in which requesters can post tasks (Human Intelligence Tasks; HITs) to be completed by workers associated with MTurk in exchange for financial compensation. MTurk has become a popular tool among researchers to collect data (Harms & Desimone, 2015). Concerns regarding the data quality of MTurk samples have arisen because participants using MTurk complete studies unsupervised, in unstructured and potentially distracting environments, and with access to external sources. Additionally, MTurk participants may be motivated to perform multiple tasks simultaneously primarily for increased financial compensation (Chandler & Shapiro, 2016). However, evaluations of the quality of the data collected through MTurk suggest that the data are of acceptable quality and meet psychometric standards (Buhrmester, Kwang, & Gosling, 2011). Additionally, MTurk provides researchers with tools to improve data quality, such as controlled access and the ability to reject low quality work. Best practice recommendations are also widely found online (e.g., Buhrmester, 2018; Chandler & Shapiro, 2016; Johnson & Borden, 2012). We followed best practice recommendations as outlined below.

To reduce the likelihood of deviant responses, we administered three instructional manipulation checks (IMCs) to the MTurk sample. IMCs (Oppenheimer, Meyvis, & Davidenko, 2009) are often used and recommended as a method to measure inattentiveness and to disqualify

deviant responders (Johnson & Borden, 2012). Research has found that excluding participants who failed IMCs reduces statistical noise thus increasing the statistical power of analyses (Goodman, Cryder, & Cheema, 2013; Oppenheimer et al., 2009). These questions were included among the items for the two new versions of the AFRAT and among the items of the ARHQ.

Despite the positive results from IMCs, their use has been criticized by researchers as an ineffective method of evaluating the quality of data. Reasons for this include that the IMCs do not assure increased attention, may change the cognitive process of participants, and could remove non-native English speakers (Goodman et al., 2013; Hauser & Schwarz, 2016). Additionally, restricting participation to workers with a high approval rate has been deemed as effective as including IMCs (Buhrmester, Talaifar, & Gosling, 2018). Taking into consideration the criticisms of IMCs, failing them alone did not disqualify participants. Participants whose only disqualification was failing the IMCs were reviewed individually by the researchers and removed based on the researchers' judgment regarding the participant's overall performance, scores, and time taken to complete the task. Out of 320 excluded participants, 47 were removed due to failure to correctly answer the IMCs.

To minimize the likelihood of inaccurate responding, we also used several other precautions. Predefined conditions that would determine whether a task was successfully completed (i.e., minimum time and completion requirements for each section of the task, consistent responses to synonymous items in the ARHQ, and instructional manipulation checks) were established and explicit instructions were provided to workers. Participants were informed of these conditions before they agreed to participate in the study, and only those who met our minimum indicators of quality work received compensation. Data collected from participants whose work was rejected were excluded from the study.

Representativeness of the MTurk sample was also a concern as MTurk workers tend to be younger and have higher education levels than the general population (Shapiro, Chandler, & Mueller, 2013). To produce a sample that would be applicable to the Air Force, we needed to match the characteristics of a military population as closely as possible. To identify participants with demographics (e.g., age, education) similar to that of a military sample, a screening questionnaire was provided to potential participants before access to the tests was given. Additionally, participants were required to be located in the United States, to have an MTurk approval rate greater than 90% and a number of HITs approved greater than 5000 (i.e., documented evidence of good past performance).

Participants were compensated $7.00 for the completion of the instruments, which consisted of one of the new versions of the AFRAT, the old AFRAT (reading comprehension and Vocabulary subtests), the Orthographic and Phonological Choice subtests, the ARHQ, and demographic questions (317 total items, with a mean response time of 101.94 minutes).

The amount of compensation for this task was selected with the goal of ensuring that it was neither unfair nor coercive. Research has suggested that fair compensation decreases the likelihood of workers misrepresenting themselves to gain access to the study, offsets workers' motivation to perform more than one task, increases the speed of data collection, and increases engagement among workers (Chandler & Shapiro, 2016).

Data collection of the MTurk sample was done in batches. During the first batch, we did not constrict the population demographics. This was done with the intention of collecting data from MTurk participants with varying ages and education levels. For the remaining data collection stages, we restricted the education and age qualifications so that only participants who reported a high school diploma (HSD) or an equivalent education level as their highest degree earned and

ages between 18-39 years old were allowed to participate in the study. This was done to resemble the demographics of the population of BMTs we test at the APAT center.

## 3.0 RESULTS AND DISCUSSION

### 3.1 Final Sample

A total of 1,274 individuals were administered the tests. Of these, 39.8 percent (%) were female. Most of the respondents reported their race as white (71.4%) and 73.2% reported a HSD as their highest degree earned. The average age of the sample was 25.7 years with most of the respondents falling in the 17-26 years age group (62.8%). Table 2 presents demographic data categorized by sample source.

**Table 2. Descriptive Statistics of the MTurk and APAT Samples**

| | APAT | | MTurk | | Total | |
|---|---|---|---|---|---|---|
| **Sex** | **N** | **%** | **N** | **%** | **N** | **%** |
| Male | 519 | 66.9% | 246 | 49.4% | 765 | 60.1% |
| Female | 255 | 32.9% | 252 | 50.7% | 507 | 39.8% |
| Missing | 1 | 0.1% | | | | 0.1% |
| **Age** | **N** | **%** | **N** | **%** | **N** | **%** |
| 17-26 | 710 | 91.5% | 90 | 18.1% | 800 | 62.8% |
| 27-36 | 55 | 7.1% | 290 | 58.2% | 345 | 27.1% |
| 37-46 | 10 | 1.3% | 89 | 17.9% | 99 | 7.8% |
| 47-56 | | | 18 | 3.6% | 18 | 1.4% |
| > 57 | | | 11 | 2.2% | 11 | 0.9% |
| Missing | 1 | 0.1% | | | 1 | 0.1% |
| **Education** | **N** | **%** | **N** | **%** | **N** | **%** |
| High School degree | 633 | 81.6% | 299 | 60.1% | 932 | 73.2% |
| Associate degree | 75 | 9.7% | 51 | 10.2% | 126 | 9.9% |
| Bachelor's degree | 53 | 6.8% | 124 | 24.9% | 177 | 13.9% |
| Master's degree | 5 | 0.6% | 24 | 4.8% | 29 | 2.3% |
| Doctorate degree | 2 | 0.3% | | | 2 | 0.2% |
| Missing | 8 | 1.0% | | | 8 | 0.6% |
| **Ethnicity** | **N** | **%** | **N** | **%** | **N** | **%** |
| Hispanic or Latino | 145 | 18.7% | 52 | 10.4% | 197 | 15.5% |
| Not Hispanic or Latino | 582 | 75.0% | 446 | 89.6% | 1028 | 80.7% |
| Missing | 49 | 6.3% | | | 49 | 3.9% |
| **Race** | **N** | **%** | **N** | **%** | **N** | **%** |
| American Indian or Alaskan Native | 5 | 0.6% | 7 | 1.4% | 12 | 0.9% |
| Asian | 37 | 4.8% | 26 | 5.2% | 63 | 5.0% |
| Black or African American | 95 | 12.2% | 58 | 11.7% | 153 | 12.0% |
| Native Hawaiian or other Pacific Islander | 7 | 0.9% | 2 | 0.4% | 9 | 0.7% |
| White | 526 | 67.8% | 383 | 76.9% | 909 | 71.2% |
| More than one | 39 | 5.0% | 22 | 4.4% | 61 | 4.8% |
| Missing | 67 | 8.6% | | | 67 | 5.3% |
| **Total** | **776** | **100%** | **498** | **100%** | **1274** | **100%** |

Three hundred and nine individuals from the Air Force received the Orthographic Choice test, Phonological Choice subtest and the ARHQ. The majority of the participants reported themselves as male (82.2%), between the ages of 17-26 (95.5%), and white (73.5%). Complete demographics for the groups are presented in Table 3.

Eight hundred and twenty-four MTurk workers attempted to complete the task. Of those, 498 completed it successfully and 320 participants were removed from the subject pool due to failure to meet predetermined indicators of quality work. Of those who successfully completed the

measures, 252 were female and 246 were male. Most reported their race as white (73.5%). Participant age ranged from 20 to 76 years, with a mean age of 33 years. Demographic data for the MTurk sample are displayed in Table 2. A roughly equal number of participants from each sample took each form of the new AFRAT (see Table 4).

**Table 3. Ortho-Phono, and ARHQ Demographic Statistics-APAT Sample**

| **Sex** | **N** | **%** |
|---|---|---|
| Male | 254 | 82.2% |
| Female | 54 | 17.5% |
| Missing | 1 | 0.3% |
| **Age** | **N** | **%** |
| 17-26 | 295 | 95.5% |
| 27-36 | 12 | 3.9% |
| 37-46 | 2 | 0.7% |
| **Education** | **N** | **%** |
| High School Diploma | 266 | 86.1% |
| Associate degree | 17 | 5.5% |
| Bachelor's degree | 21 | 6.8% |
| Master's degree | 0 | 0.0% |
| Doctorate degree | 1 | 0.3% |
| Missing | 4 | 1.3% |
| **Ethnicity** | **N** | **%** |
| Hispanic or Latino | 49 | 15.7% |
| Not Hispanic or Latino | 234 | 75.7% |
| Missing | 26 | 8.4% |
| **Race** | **N** | **%** |
| American Indian or Alaskan Native | 3 | 1.0% |
| Asian | 13 | 4.2% |
| Black or African American | 23 | 7.4% |
| Native Hawaiian or other Pacific Islander | 2 | 0.7% |
| White | 227 | 73.5% |
| More than one | 19 | 6.2% |
| Missing | 22 | 7.1% |

**Table 4. Frequencies of Individuals who Took Each New Form**

| | **APAT** | **MTurk** | **Total** |
|---|---|---|---|
| Form A | 401 | 240 | 641 |
| Form B | 375 | 258 | 633 |
| Total | 776 | 498 | 1274 |

*Note:* Examinees in this sample were given either Form A or Form B, and AFRAT Form 81A (reading comprehension and vocabulary)

## 3.2 Final Tests

### 3.2.1. Reading Comprehension

The new reading comprehension forms each contain 40 items drawn from seven passages. The test has a time limit of 35 minutes. The test is scored by adding the number of correct responses with no penalty incurred for skipping items or answering them incorrectly. The data collected from MTurk closely matched the APAT data despite the different administration methods (i.e., computerized vs paper-and-pencil). This suggests that if the AFRAT were to be computerized in the future, the examinees' scores would likely not differ substantially from the paper-and-pencil version. Despite limited capacity to monitor testing fidelity in the MTurk sample (e.g., we could not force close all other browsing windows), the data show similar means, standard deviations, and median scores as the proctored APAT sample. However, Form 81A had a slightly higher mean than Form A and Form B. The MTurk sample performed slightly better in Form 81A than the APAT sample. Table 5 and Table 6 show the summary statistics from each sample for all test versions.

**Table 5. Form 81A Summary Statistics**

| | Form 81A | | |
|---|---|---|---|
| | **Combined** | **APAT** | **MTurk** |
| N | 1274 | 776 | 498 |
| Mean | 33.35 | 32.75 | 34.28 |
| SD | 5.72 | 5.85 | 5.39 |
| Median | 35 | 34 | 36 |
| Skewness | -1.69 | -1.49 | -2.1 |
| Kurtosis | 3.32 | 2.53 | 5.39 |

**Table 6. Reading Comprehension Summary Statistics for Forms A and B**

| | Form A | | | Form B | | | |
|---|---|---|---|---|---|---|---|
| | **Combined** | **APAT** | **MTurk** | **Combined** | **APAT** | **MTurk** | |
| N | 641 | 401 | 240 | 633 | 375 | 258 | |
| Mean | 31.65 | 31.70 | 31.57 | 31.71 | 31.69 | 31.73 | |
| SD | 6.30 | 6.30 | 6.32 | 5.75 | 5.63 | 5.93 | |
| Median | 33 | 33 | 33 | 33 | 33 | 33.5 | |
| Skewness | -1.27 | -1.25 | -1.31 | -1.27 | -1.04 | -1.58 | |
| Kurtosis | 1.70 | 1.37 | 2.31 | 1.73 | 0.92 | 2.72 | |

As can be seen in Table 6, the descriptive statistics of the two new forms match closely. Average item difficulty was .80 for both forms and the average biserial correlation was .57 for both versions.

Item difficulty statistics for the reading comprehension tests are presented in Table 7. The average difficulty of the new forms was .80 compared to .84 for Form 81A. The average difficulty (.80) was the same in both new forms, and the biserial correlation was slightly higher for Form A (.59) than Form B (.55). Table 7 and Table 8 show the breakdown of item difficulties and biserial correlations, respectively. A breakdown of item difficulties and biserial correlations from each sample can be found in Appendix B. Most items from Form A (83%) and Form B (73%) fall within a difficulty range of .70-.89, in contrast to only 50% of items falling in this range for Form 81A. Therefore, the new forms had a wider range of difficulties, which meets the goal of preventing items from being so easy that almost all examinees answer them correctly regardless of RGL.

**Table 7. Item Difficulty Summary**

| P-Value | 81 A (Old AFRAT) | Form A | Form B |
|---|---|---|---|
| .90-.99 | 15 | 2 | 7 |
| .80-.89 | 16 | 21 | 18 |
| .70-.79 | 4 | 12 | 11 |
| .60-.69 | 3 | 5 | 2 |
| .59 > | 2 | 0 | 2 |
| Average p-value | .84 | .80 | .80 |

**Table 8. Biserial Correlation Summary**

| Biserial | 81 A (Old AFRAT) | Form A | Form B |
|---|---|---|---|
| .70-.99 | 20 | 10 | 9 |
| .50-.69 | 15 | 18 | 16 |
| .30-.49 | 5 | 12 | 13 |
| .29 > | 0 | 0 | 2 |
| Average biserial | .70 | .59 | .55 |

Analyses of internal consistency reliability (Kuder-Richardson, Formula 20) were conducted for the combined samples. Overall, the new forms had good internal reliability. The

average internal consistency reliability for the new versions of the AFRAT was .84. Form A had a reliability of .86, and form B had a reliability of .83. Compared to Form 81A (.88), the new versions had slightly lower reliability.

The new versions of the AFRAT correlated well with Form 81A (Form A, $r$ = .75; Form B, $r$ = .80). Table 9 shows the correlations among the instruments.

**Table 9. Test Correlations**

| | Form A | Form B | Form 81A | Vocab | ARHQ | Ortho | Phono |
|---|---|---|---|---|---|---|---|
| Form A | 1 | | | | | | |
| Form B | - | 1 | | | | | |
| Form 81A | 0.75** | 0.80** | 1 | | | | |
| Vocabulary | 0.63*** | 0.63** | 0.66** | 1 | | | |
| ARHQ | -0.20** | -0.18** | -0.24** | -0.26** | 1 | | |
| Orthographic | 0.28** | 0.27** | 0.22** | 0.33** | -0.28** | 1 | |
| Phonological | 0.35** | 0.40** | 0.37** | 0.37** | -0.19** | 0.39** | 1 |

** Correlation significant at $p$ <.001

### 3.2.2. Orthographic and Phonological Choice Subtests

The Phonological and Orthographic Choice subtests each contain 60 items. Each subtest has a time limit of two minutes. Each subtest is scored by taking the total number of correct answers and subtracting the total number of incorrect or skipped items up to the last item completed. Given the speeded nature of the tests, examinees were not expected to answer the 60 items within the time limit.

The Orthographic Choice subtest, with an average p-value of .88, was substantially less difficult than the Phonological Choice subtest which had an average p-value of .78. Summary statistics for the Orthographic and Phonological Choice subtests are in Table 10 and their p-value distributions are in Table 11.

Concerning the usefulness of using MTurk participants to guide future use of their data by the Air Force, we identified differences in test performance between Airmen and MTurk examinees. The Phonological Choice subtest means were essentially the same for both groups (26.2 for Airmen vs 25.6 for MTurk). The Orthographic Choice subtest was notably easier for MTurk examinees (Mean = 34.6, SD = 13.01) than for Airmen examinees (Mean = 31.0, SD = 13.96), with a mean p-value of .84 for the Air Force sample and a mean p-value of .93 for MTurk.

The MTurk/Airmen differences may be attributed to the format of the subtest (i.e., paper-and-pencil vs computer-based). Additionally, differences in age, education levels, and test setting between the samples may have contributed to these differences.

Significant gender differences were observed in the Orthographic subtest, $t(649.33)=3.67$, $p$ <.001[1], with females performing better (M = 35.47, SD = 13.31) than males (M = 31.91, SD = 13.43). There was no significant gender difference for the Phonologic subtest. This aligns with the findings of past research (i.e., Wolff & Lundberg, 2003).

[1] All $t$-tests reported were Welch's unequal variances $t$-tests.

**Table 10. Orthographic-Phonological Choice Summary Statistics**

| | Orthographic | | | Phonological | | |
|---|---|---|---|---|---|---|
| | **Combined** | **APAT** | **MTurk** | **Combined** | **APAT** | **MTurk** |
| N | 807 | 309 | 498 | 807 | 309 | 498 |
| Mean | 33.26 | 31.04 | 34.64 | 25.81 | 26.21 | 25.56 |
| SD | 13.49 | 13.96 | 13.01 | 14.01 | 13.67 | 14.22 |
| Median | 33 | 31 | 34 | 28 | 28 | 28 |
| Skewness | -0.38 | -0.63 | -0.16 | -1.03 | -0.67 | -1.23 |
| Kurtosis | 1.01 | 2.13 | -0.17 | 1.29 | 0.02 | 1.89 |

**Table 11. Orthographic-Phonological Choice p-values**

| | Orthographic | | Phonological | |
|---|---|---|---|---|
| **Difficulty** | **APAT** | **MTurk** | **APAT** | **MTurk** |
| .90-.99 | 23 | 43 | 12 | 18 |
| .80 - .89 | 18 | 17 | 22 | 19 |
| .70 - .79 | 14 | 0 | 16 | 10 |
| .60 - .69 | 2 | 0 | 7 | 3 |
| .59 and less | 3 | 0 | 3 | 10 |
| Total Items | 60 | 60 | 60 | 60 |
| Average | 0.84 | 0.93 | 0.80 | 0.78 |
| Combined | 0.88 | | 0.78 | |

## 3.3 Validity

Participants who took the ARHQ were categorized into one of two groups based on a cutoff score of .4, derived by Lefly and Pennington (2000). Those whose score was greater than or equal to .4 were categorized as being at risk of a reading disability and those below the .4 cutoff score were categorized as having no risk. The mean score of the ARHQ was .27 (SD = 0.11). From the total sample, 10.16% were categorized as being at risk of a reading disability. The modified ARHQ showed good reliability (Cronbach's alpha = .78).

The Air Force sample scored significantly higher ($p$<.001) in the ARHQ (M = 0.31, SD = 0.06) than the MTurk sample (M = 0.24, SD = 0.13). However, in the Air Force sample only 6.1% were at risk, whereas 12.8% of the MTurk sample was at risk.

Males (M = 0.28, SD = 0.1) scored significantly higher ($p$ < .001) than females (M = 0.24, SD = 0.13), In the female group, 12.75% were categorized as being at risk, while in the male group 8.6% were categorized as being at risk.

The apparent contradiction between one group scoring higher on the ARHQ but having a lower percentage at risk can be understood by examining Figures 1 and Figure 2 which show the distribution of the ARHQ split by sample source and gender respectively. The MTurk distribution has a larger range and is more unevenly distributed than the Air Force distribution, which is more symmetrical and with most scores falling between .2 and .4. Similarly, in the male distribution, most scores fall between .2 and .4, and the female distribution is flatter than the male distribution.

The Orthographic Choice subtest had a notably stronger correlation with the ARHQ ($r$ = -.28) than did the Phonological Choice subtest ($r$= -.19). The ARHQ had similar correlations with the new forms of the AFRAT (see Table 9).

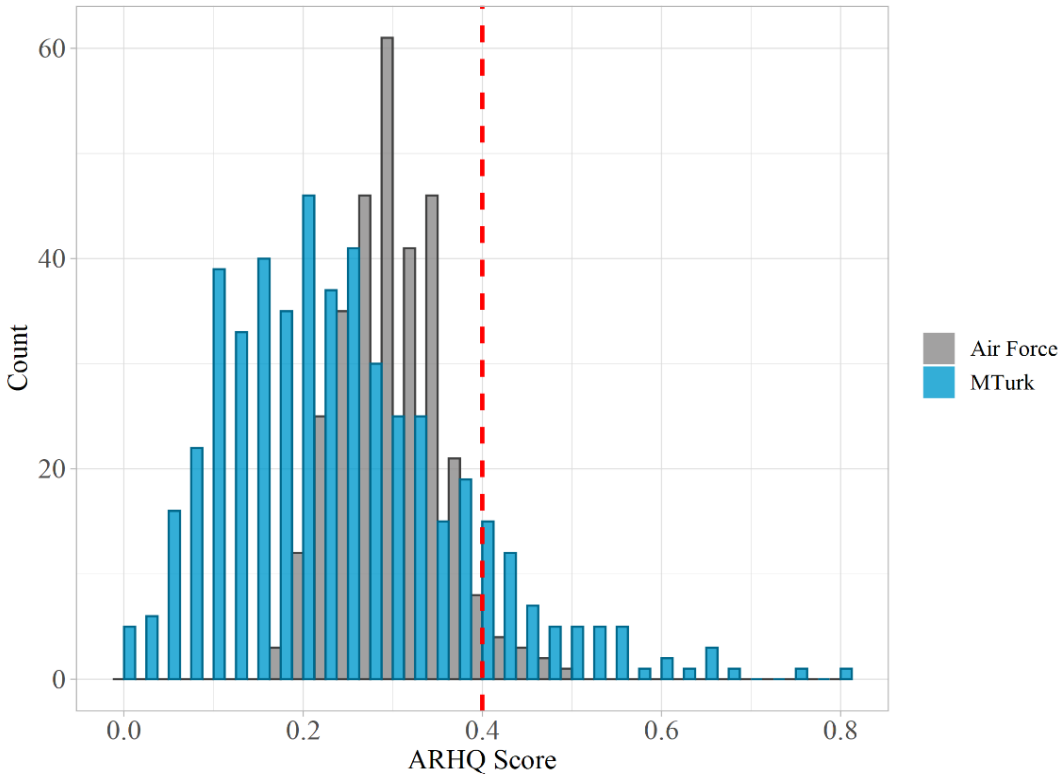



**Figure 1. ARHQ Distribution Split by Sample**

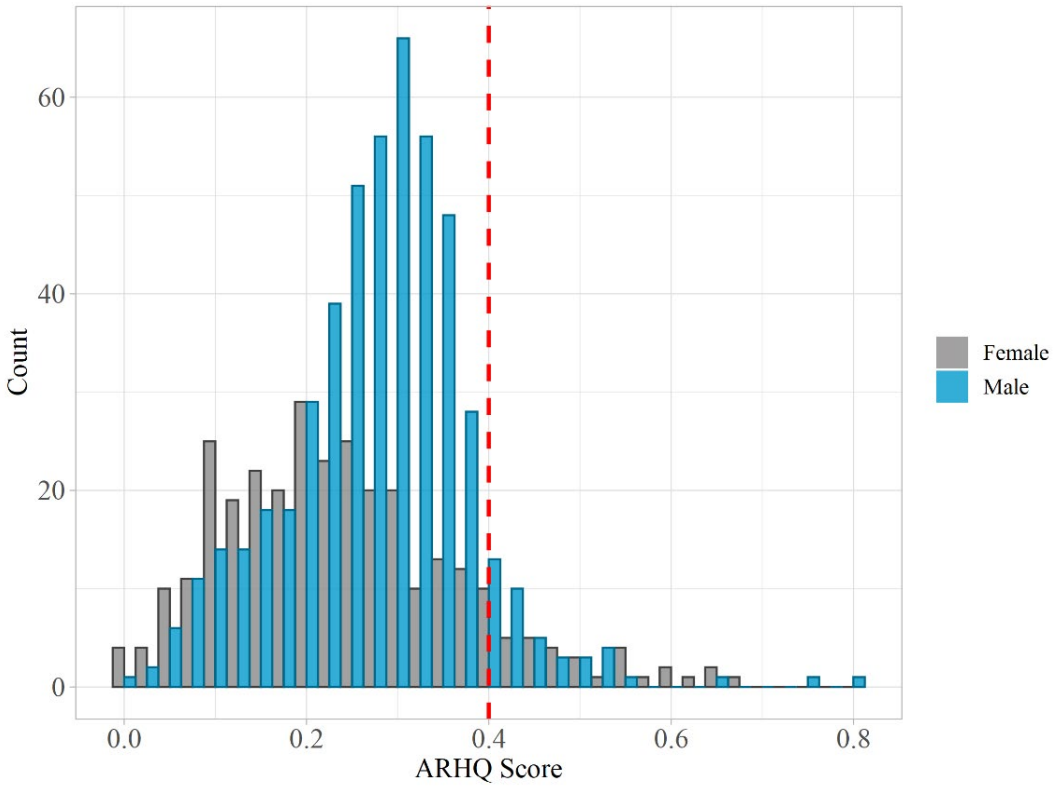



**Figure 2. ARHQ Distribution Split by Gender**

### 3.4 Equating

Equating is the process of determining scores that are equivalent to the scores on a different form (Livingston, 2014; Price, 2017). Equating enables test administrators to have a common metric and provide consistent meaning for different versions of a test (Ree, Mathews, Mullins, & Massey, 1982). Raw score to scale equipercentile equating was conducted to equate the composite New AFRAT scores to AFRAT 81A composite scores and its corresponding RGL. Composite New AFRAT scores were calculated by adding the scores of the vocabulary and reading comprehension subtests.

In equipercentile equating, scores from two forms are equivalent if they have the same percentile rank (Angoff, 1984). To equate two scores with the same percentile rank, the scores in the new form were transformed to the scores on the reference forms. In the cases where the same percentile rank is not found on the reference form, interpolation was used to adjust the scores.

Raw score to RGL conversion tables are presented in Appendix C, figures depicting score distributions and equating comparisons are shown in Appendix D. Summary statistics using the equated scores show that the converted scores have roughly the same skewness and kurtosis as Form 81A (Table 12). Appendix E presents a RGL conversion tables based on a second equating method.

**Table 12. Summary Statistics of Equated Scores**

| | Mean | SD | Skewness | Kurtosis |
|---|---|---|---|---|
| Form 81A Composite | 70.1 | 10.6 | -1.4 | 2.3 |
| Form A Equated Composite | 70.1 | 10.7 | -1.4 | 2.4 |
| Form B Equated Composite | 70.1 | 10.7 | -1.4 | 2.5 |

## 3.5 Orthographic and Phonological Choice Use

Three measures that can contribute to a determination that an Airman is at risk were examined in this study: RGL, Orthographic, and Phonological Choice subtest scores. Given the limitations of the available data (i.e., obtaining data from separate population pools) and the lack of a strong criterion, we make the following suggestions for operational use.

We created two methods of combining the three measures. These methods are based on a user inputting a threshold for each of the measures. The threshold returns a percentage of individuals in the test sample identified as being at risk. The first method identifies Airmen who meet all conditions (e.g., the user can determine the percentage of Airmen who are in the bottom 25$^{th}$ percentile for each test and below a 10 RGL; this would identify 5.8% of individuals in our sample as being at risk). The output would indicate what percentage of Airmen meet those conditions. Table 13 shows a sample output.

The second method identifies those with a RGL below the threshold established, and either below the Orthographic threshold or the Phonological threshold. This method identifies individuals who meet two of the conditions. Table 14 shows a sample of the input and output from that approach.

**Table 13. User Inputs and Result with All Conditions Met**

| Measure | Inputs[1] | Inputs[2] |
|---|---|---|
| Orthographic percentile | 0.25 | 0.20 |
| Phonological percentile | 0.25 | 0.20 |
| Reading Grade Level | 10.0 | 10.5 |
| **% At Risk** | **5.8%** | **5.7%** |

Note: Inputs identify individuals below the specified percentile.
[1] Example 1. [2] Example 2

**Table 14. User Inputs and Result with Two Conditions Met**

| Measure | Inputs[1] | Inputs[2] |
|---|---|---|
| Orthographic percentile | 0.10 | 0.20 |
| Phonological percentile | 0.10 | 0.20 |
| Reading Grade Level | 10.0 | 9.5 |
| **% At Risk** | **7.9%** | **6.6%** |

Note: Inputs identify individuals below the specified percentile.
Gray boxes indicate the condition not met.
[1] Example 1. [2] Example 2

The user can determine inputs based on an absolute threshold of Airmen performance, such as below a given percentile for the tests, or the user can adjust the thresholds to achieve a target percentage of Airmen identified as at risk. These tables will be more accurate for Airmen as the database of scores grows larger with data only from Airmen. A provisional template developed to develop these tables is shown in Appendix F. Appendix G shows graphs representing the relationship of the Orthographic and Phonological Choice tests and RGL.

## 4.0 CONCLUSIONS AND RECOMMENDATIONS

The new reading comprehension forms represent an improvement in content over the previous versions and meet high psychometric standards. The new forms are parallel and have been calibrated to the old form and the accompanying RGL scale. The new forms now include content that is relevant to the context of enlisted Air Force servicemembers and have a slightly improved discriminatory ability.

Two subtests have been added to the AFRAT: the Orthographic and Phonological Choice subtests. They are designed to identify Airmen with potential reading disabilities. Examinations of the psychometric properties of the new subtests provided evidence of good reliability and validity.

Since the data collected in this study depended on the use of test subjects who were not members of the Air Force, we recommend that a larger sample of Airmen be tested to provide a more accurate estimate of the test metrics. The two additional sections and the RGL measure can be used in combination to identify Airmen who fail to meet threshold performance. Alternatively, the three measures can be used to identify a given percentage of Airmen who are most at risk.

In conclusion, the updated AFRAT and the two new subtests are an effective tool for identifying Air Force members with potential reading disabilities, allowing for referral to additional testing or remediation.

## 5.0 REFERENCES


American Speech-Language-Hearing Association. (n.d.). *Disorders of reading and writing.* https://www.asha.org/Practice-Portal/Clinical-Topics/Written-Language-Disorders/Disorders-of-Reading-and-Writing/

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

Bruck, M. (1992). Persistence of dyslexics' phonological awareness deficits. *Developmental Psychology, 28*(5), 874-886.

Buhrmester, M. D. (2018, October). *Amazon Mechanical Turk guide for social scientists.* Retrieved February, 5, 2020, from https://michaelbuhrmester.com/mechanical-turk-guide/

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*(2), 149-154.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, *6*(1), 3-5. https://doi.org/10.1177/1745691610393980

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual review of clinical psychology*, *12*.

Fowler, A. E., & Scarborough, H. S. (1993). *Should reading-disabled adults be distinguished from other adults seeking literacy instruction?* (Report No. 93-7). Philadelphia, PA: National Center on Adult Literacy.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213-224.

Grigorenko, E. L. (2001). Developmental dyslexia: An update on genes, brains, and environments. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 42,* 91-125. https://doi.org/10.1017/S0021963001006564

Harms, P. D., & DeSimone, J. A. (2015). Caution! MTurk workers ahead--Fines doubled. *Industrial and Organizational Psychology*, *8*(2), 183-190. https://doi.org/10.1017/iop.2015.23

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, *48*(1), 400-407.

Johnson, D. R., & Borden, L. A. (2012). Participants at your fingertips: Using Amazon's Mechanical Turk to increase student–faculty collaborative research. *Teaching of Psychology*, *39*(4), 245-251.

Kemp, N., Parrila, R. K., & Kirby, J. R. (2008). Phonological and orthographic spelling in high-functioning adult dyslexics. *Dyslexia, 15,* 105-128. https://doi.org/10.1002/dys.364

Landi, N., & Ryherd, K. (2017). Understanding specific reading comprehension deficit: A review. *Language and Linguistic Compass, 11*(2).

Lefly, D. L., & Pennington, B. F. (2000). Reliability and validity of the adult reading history questionnaire. *Journal of Learning Disabilities*, *33*(3), 286-296.

Livingston, S. A. (2014). Equating test scores (without IRT). *Educational testing service*.

Massey, R. H., & Mathews, J. J. *Reading grade levels of Air Force civilian personnel.* AFHRL-TR-80-11, AD-A087066, Brooks AFB, TX: Manpower and Personnel Division, July 1980.

Mathews, J. J., & Roach, B. W. (1983). *Reading abilities tests: Development and norming for Air Force use.* AFHRL-TR-82-26, AD-A125 913. Brooks AFB, TX: Manpower and Personnel Division.

Mathews, J. J., Valentine, L. D., Jr., & Sellman, W. S. (1978). *Prediction of reading grade levels of service applicants from Armed Services Vocational Aptitude Battery (ASVAB).* AFHRL-TR-78-82, AD-A063 656. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Mortimore, T., & Crozier, W. R. (2006). Dyslexia and difficulties with study skills in higher education. *Studies in Higher Education, 31*(2), 235-251. https://doi.org/10.1080/03075070600572173

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867-872.

Price, L. R. (2016). *Psychometric methods: Theory into practice*. Guilford Publications.

Ree, M. J., Mathews, J. J., Mullins, C .J., Massey, R. H. (1982). *Calibration of Armed Services Vocational Aptitude Battery Forms 8, 9, and 10.* AFHRL-TR-81-49, AD-A114 714, Brooks AFB, TX, Air Force Human Resources Laboratory.

Riemer, S. E. (1984). *Air force reading abilities test: Utilization assessment.* AFHRL-SR-83-23, AD-A138 986. Brooks AFB, TX: Applications and Liaison Office.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213-220.

Shaywitz, S. E. (1998). Dyslexia. *The New England Journal of Medicine, 338,* 307-312. https://doi.org/10.1056/NEJM199801293380507

Shaywitz, S. E., Gruen, J. R., & Shaywitz, B. A. (2007). Management of dyslexia, its rationale, and underlying neurobiology. *Pediatric Clinics of North America, 54*, 609-623. https://doi.org/doi:10.1016/j.pcl.2007.02.013

Skinner, J., Thompson, N., Schwartz, K., & Weissmuller, J. (2007). *Air Force personnel research issues: A manager's handbook* (AFCAPS-FR-2010-0017). Randolph AFB, TX: Air Force Personnel Center – Strategic Research and Analysis Branch.

Treiman, R. (1997). Spelling in normal children and dyslexics. In B. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention*. Mahwah, NJ: Lawrence Erlbaum Associates.

Vellutino, F. R. (1987). Dyslexia. *Scientific American, 256*(3), 34-41. Retrieved from https://www.jstor.org/stable/24979338

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry, 45*(1), 2-40.

Wanzek, J., Wexler, J., Vaughn, S., & Ciullo, S. (2010). Reading interventions for struggling readers in the upper elementary grades: A synthesis of 20 years of research. *PubMed Central, 23*(8), 889-912. https://doi.org/10.1007/s11145-009-9179-5

Welcome, S. E., & Meza, R. A. (2019). Dimensions of the Adult Reading History Questionnaire and their relationships with reading ability. *Reading and Writing*, *32*(5), 1295-1317.

Wolff, U., & Lundberg, I. (2003). A technique for group screening of dyslexia among adults. *Annals of Dyslexia, 53,* 324-339.

Woodruff, S., Schumaker, J. B., & Deshler, D. D. (2002). *The effects of an intensive reading intervention on the decoding skills of high school students with reading deficits* (Report No. 15). Washington, DC: Special Education Programs.

**APPENDIX A – The Adult Reading History Questionnaire**

## Part 7
## Reading Preferences Questionnaire

**DIRECTIONS:** This part of the test will ask you questions about your experience and preference towards reading. There are not *right* or *wrong* answers – the goal is to record your preferences and experience. Read each statement and, based on your experience, choose the one that best describes you and mark the bubble that corresponds in the answer sheet.

You should work quickly and respond to all statements. Do not spend too much time deciding what your answer should be; just answer based on your general experience with reading. Please answer each question honestly by selecting the option that best describes your attitude or experience for each statement.
This part of the test will begin in **Part X**, Question X and end on Question X.
When you complete all statements, place your booklets at the front edge of your desk so the proctor will know that you are done.

**Please respond to the following questions honestly.**

1. Is English your first language?
   A) Yes B) No
2. Have your parents or siblings been diagnosed with a reading disability (e.g. Dyslexia)?
   A) Yes B) No
3. Have you ever been diagnosed with a reading disability (e.g. Dyslexia)?
   A) Yes
   B) No

**Please select the response that most nearly describes your attitude or experience for each of the following questions or statements.**

4. * How would you compare your current reading speed to that of others of the same age and education?
   A) Below Average
   B) Slightly Below Average
   C) Average
   D) Slightly Above Average
   E) Above Average

5. * How much reading do you do in conjunction with your work (if not working, how much did you read when you were working?
   A) None
   B) Little
   C) Some
   D) Much
   E) A Great Deal

6. How much difficulty did you have learning to spell in elementary school?
   A) None
   B) Little
   C) Some
   D) Much
   E) A Great Deal

7. * How would you compare your current spelling to that of others of the same age and education?
   A) Below Average
   B) Slightly Below Average
   C) Average
   D) Slightly Above Average
   E) Above Average

8. Did your parents ever consider having you repeat any grades in school due to academic failure (not illness)?
   A) No
   B) Talked about it, but did not do it
   C) Repeated 1 grade
   D) Repeated 2 grades
   E) Dropped out

9. Do you ever have difficulty remembering people's names or names of places?
   A) No
   B) A Little
   C) Some
   D) Much
   E) A Great Deal

10. Do you have difficulty remembering addresses, phone numbers, or dates?
    A) None
    B) A Little
    C) Some
    D) Much
    E) A Great Deal

11. Do you have difficulty remembering complex verbal instructions?
    A) No
    B) A Little
    C) Some
    D) Much
    E) A Great Deal

12. Do you currently reverse the order of letters or numbers when you read or write?
    A) No
    B) A Little
    C) Some
    D) Much
    E) A Great Deal

13. * How many books or eBooks do you read for pleasure each year?
    A) More than 10
    B) 6-10
    C) 2-5
    D) 1-2
    E) None

14. * Which of the following most nearly describes your attitude toward school when you were a child:
    A) Hated School; tried to get out of going
    B) Disliked School
    C) Neither liked nor disliked
    D) Enjoyed it
    E) Loved it; favorite activity

15. How much difficulty did you have learning to read in elementary school?
    A) None
    B) A Little
    C) Some
    D) Much
    E) A Great Deal

16. How much extra help did you need when learning to read in elementary school?
    A) No help
    B) Help from others
    C) Help from teachers/parents
    D) Tutors or special class (1 year)

E) Tutors or special class (2+ years)

17. Did you ever reverse the order of letters or numbers when you were a child?
A) No
B) Rarely
C) Sometimes
D) Often
E) Very often

18. (IMC) Do you carefully read every survey item, if so, select A Great Deal?
A) No
B) Little
C) Some
D) Much
E) A Great Deal

19. Did you have difficulty learning letter and/or color names when you were a child?
A) No
B) A Little
C) Some
D) Much
E) A Great Deal

20. * How would you compare your reading skill to that of others in your elementary classes?
A) Below Average
B) Slightly Below Average
C) Average
D) Slightly Above Average
E) Above Average

21. All students struggle from time to time in school. Compared to others in your classes, how much did you struggle to complete your work?
A) Not at all
B) Less than most
C) About the same
D) More than most
E) Much more than most

22. Did you experience difficulty in high school or college English classes?
A) No
B) A Little
C) Some
D) Much
E) A Great Deal

23. * What is your current attitude toward reading?
    A) Very negative
    B) Slightly negative
    C) Neutral
    D) Slightly positive
    E) Positive

24. * How much reading do you do for pleasure?
    A) No
    B) A Little
    C) Some
    D) Much
    E) A Great Deal

* denotes a reverse coded item. IMC denotes an Instructional Manipulation Check

## APPENDIX B – Summary Statistics Split by Sample Size

**Table B-1. Distribution of P-Values Split by Sample**

| P-value | 81 A (Old) | | Version A | | Version B | |
|---|---|---|---|---|---|---|
| | APAT | MTurk | APAT | MTurk | APAT | MTurk |
| .90 - .99 | 13 | 23 | 2 | 6 | 10 | 7 |
| .80 - .89 | 18 | 10 | 23 | 17 | 13 | 14 |
| .70 - .79 | 2 | 4 | 10 | 11 | 13 | 13 |
| .60 -.69 | 4 | 2 | 4 | 5 | 2 | 4 |
| .59 < | 3 | 1 | 1 | 1 | 2 | 2 |

**Table B-2. Distribution of Biserial Correlations Split by Sample**

| Biserial | 81 A (Old AFRAT) | | Version A | | Version B | |
|---|---|---|---|---|---|---|
| | APAT | MTurk | APAT | MTurk | APAT | MTurk |
| .70 - .99 | 18 | 24 | 13 | 7 | 6 | 10 |
| .50 - .69 | 14 | 12 | 15 | 28 | 19 | 18 |
| .30 - .49 | 8 | 4 | 9 | 2 | 12 | 11 |
| <.29 | 0 | 0 | 3 | 2 | 3 | 1 |

## APPENDIX C – AFRAT Form A and B Conversion Tables

**Table C-1. Reading Grade Level Conversion for Form A**

| Raw Score | RGL | Raw Score | RGL |
|---|---|---|---|
| 13 | 6.9 | 50 | 8.6 |
| 14 | 6.9 | 51 | 8.8 |
| 15 | 6.9 | 52 | 8.9 |
| 16 | 6.9 | 53 | 9 |
| 17 | 6.9 | 54 | 9.2 |
| 18 | 6.9 | 55 | 9.3 |
| 19 | 6.9 | 56 | 9.4 |
| 20 | 6.9 | 57 | 9.5 |
| 21 | 6.9 | 58 | 9.6 |
| 22 | 7.1 | 59 | 9.7 |
| 23 | 7.1 | 60 | 9.7 |
| 24 | 7.1 | 61 | 9.8 |
| 25 | 7.2 | 62 | 10 |
| 26 | 7.2 | 63 | 10.1 |
| 27 | 7.3 | 64 | 10.3 |
| 28 | 7.4 | 65 | 10.5 |
| 29 | 7.7 | 66 | 10.6 |
| 30 | 7.7 | 67 | 10.7 |
| 31 | 7.8 | 68 | 10.9 |
| 32 | 7.8 | 69 | 11.1 |
| 33 | 7.7 | 70 | 11.2 |
| 34 | 7.8 | 71 | 11.4 |
| 35 | 7.9 | 72 | 11.5 |
| 36 | 7.9 | 73 | 11.7 |
| 37 | 7.9 | 74 | 11.9 |
| 38 | 7.9 | 75 | 12.2 |
| 39 | 7.9 | 76 | 12.3 |
| 40 | 8 | 77 | 12.4 |
| 41 | 8 | 78 | 12.4 |
| 42 | 8.1 | 79 | 12.5 |
| 43 | 8.1 | 80 | 12.6 |
| 44 | 8.2 | 81 | 12.7 |
| 45 | 8.3 | 82 | 12.9 |
| 46 | 8.5 | 83 | 12.9 |
| 47 | 8.5 | 84 | 12.9 |
| 48 | 8.5 | 85 | 12.9 |
| 49 | 8.6 | | |

**Table C-2. Reading Grade Level Conversion for Form B**

| Raw Score | RGL | Raw Score | RGL |
|---|---|---|---|
| 26 | 6.9 | 56 | 9.2 |
| 27 | 6.9 | 57 | 9.3 |
| 28 | 6.9 | 58 | 9.4 |
| 29 | 6.9 | 59 | 9.5 |
| 30 | 6.9 | 60 | 9.6 |
| 31 | 7.1 | 61 | 9.7 |
| 32 | 7.2 | 62 | 9.9 |
| 33 | 7.2 | 63 | 10 |
| 34 | 7.2 | 64 | 10.2 |
| 35 | 7.2 | 65 | 10.4 |
| 36 | 7.3 | 66 | 10.6 |
| 37 | 7.4 | 67 | 10.7 |
| 38 | 7.8 | 68 | 10.9 |
| 39 | 7.8 | 69 | 11 |
| 40 | 7.8 | 70 | 11.1 |
| 41 | 7.9 | 71 | 11.3 |
| 42 | 8 | 72 | 11.5 |
| 43 | 8 | 73 | 11.6 |
| 44 | 8.1 | 74 | 11.8 |
| 45 | 8.2 | 75 | 12.1 |
| 46 | 8.3 | 76 | 12.3 |
| 47 | 8.4 | 77 | 12.4 |
| 48 | 8.5 | 78 | 12.4 |
| 49 | 8.6 | 79 | 12.5 |
| 50 | 8.7 | 80 | 12.6 |
| 51 | 8.8 | 81 | 12.7 |
| 52 | 8.9 | 82 | 12.9 |
| 53 | 9 | 83 | 12.9 |
| 54 | 9.1 | 84 | 12.9 |
| 55 | 9.2 | 85 | 12.9 |

## APPENDIX D – Equating Distribution Graphs

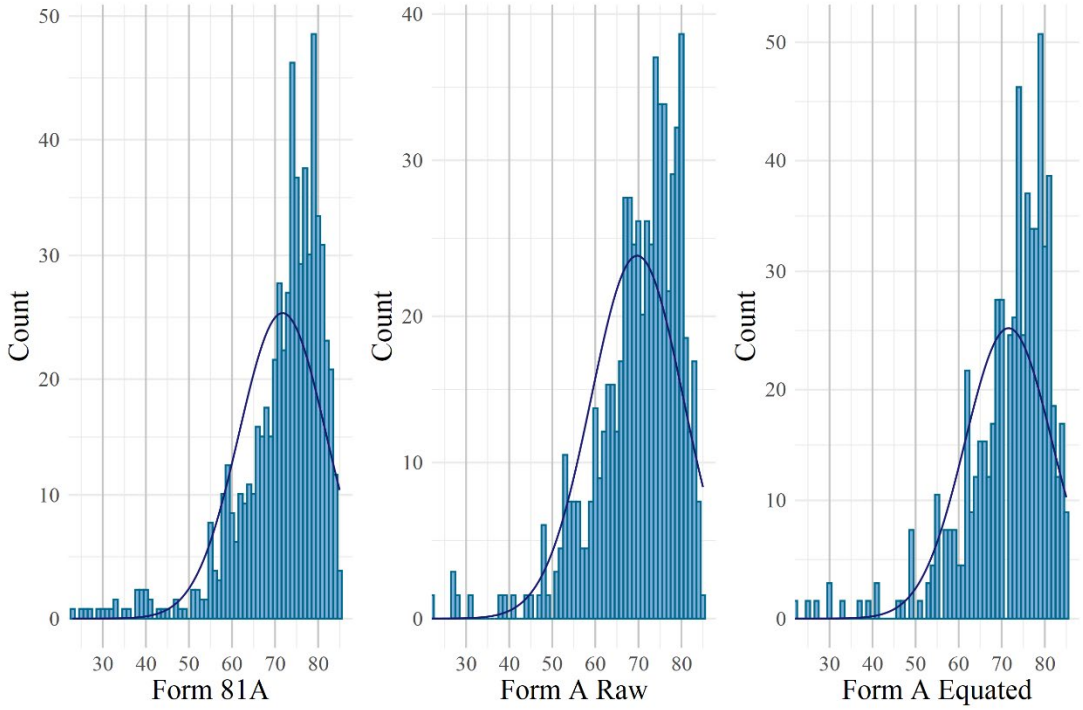



**Figure D-1. Score Distribution of Form A**

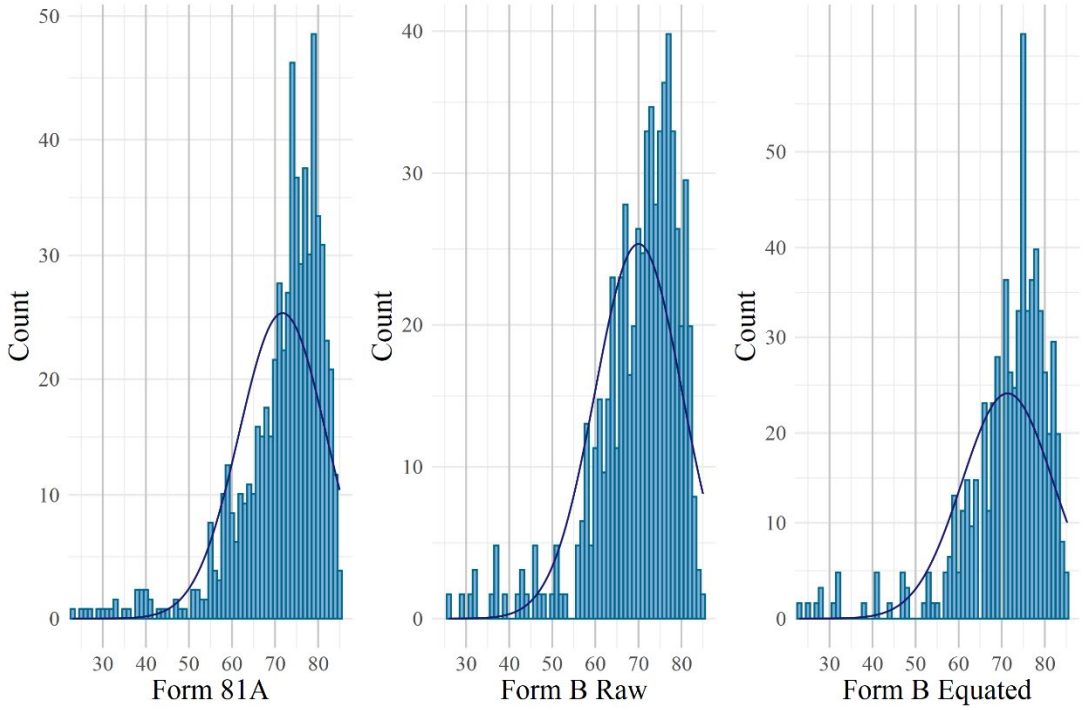



**Figure D-2. Score Distribution of Form B**

## APPENDIX E – Second Equating Method

OpTech explored different methods for equating the new forms, they all yielded very similar results and the best method was selected. Below is another equating method we explored but did not select it.

For each of the new versions of the AFRAT, there were two steps in determining each score's RGL. The RGL was known for scores on the old version. The first step was to use equipercentile equating to establish the percentile on the new version so that it matched the percentile for the old version. The RGL is known for the old version, so the next step was to attach a RGL to the new score based on its equipercentile matched item on the old test. In keeping with current practice, RGLs were rounded to one decimal. If the computation resulted in a two-decimal number where the second place was a five (i.e., .X5), the number was rounded down, very slightly favoring the identification of a problem.

The process was repeated for the second new version. The results for Form A are shown in Table E-1, and for Form B in Table E-2.

**Table E-1 Reading Grade Level for Form A**

| Score | RGL | Score | RGL |
|---|---|---|---|
| 26 | 7.1 | 56 | 9.4 |
| 27 | 7.2 | 57 | 9.4 |
| 28 | 7.3 | 58 | 9.5 |
| 29 | 7.3 | 59 | 9.6 |
| 30 | 7.3 | 60 | 9.7 |
| 31 | 7.5 | 61 | 9.7 |
| 32 | 7.8 | 62 | 9.8 |
| 33 | 7.8 | 63 | 10 |
| 34 | 7.8 | 64 | 10.2 |
| 35 | 7.8 | 65 | 10.4 |
| 36 | 7.9 | 66 | 10.5 |
| 37 | 7.9 | 67 | 10.6 |
| 38 | 7.9 | 68 | 10.7 |
| 39 | 7.9 | 69 | 10.9 |
| 40 | 8 | 70 | 11 |
| 41 | 8 | 71 | 11.1 |
| 42 | 8.1 | 72 | 11.3 |
| 43 | 8.2 | 73 | 11.5 |
| 44 | 8.3 | 74 | 11.6 |
| 45 | 8.4 | 75 | 12 |
| 46 | 8.5 | 76 | 12.2 |
| 47 | 8.6 | 77 | 12.3 |
| 48 | 8.6 | 78 | 12.4 |
| 49 | 8.8 | 79 | 12.4 |
| 50 | 8.9 | 80 | 12.5 |
| 51 | 9 | 81 | 12.6 |
| 52 | 9 | 82 | 12.7 |
| 53 | 9.1 | 83 | 12.9 |
| 54 | 9.2 | 84 | 12.9 |
| 55 | 9.3 | 85 | 12.9 |

**Table E-2 Reading Grade Level for Form B**

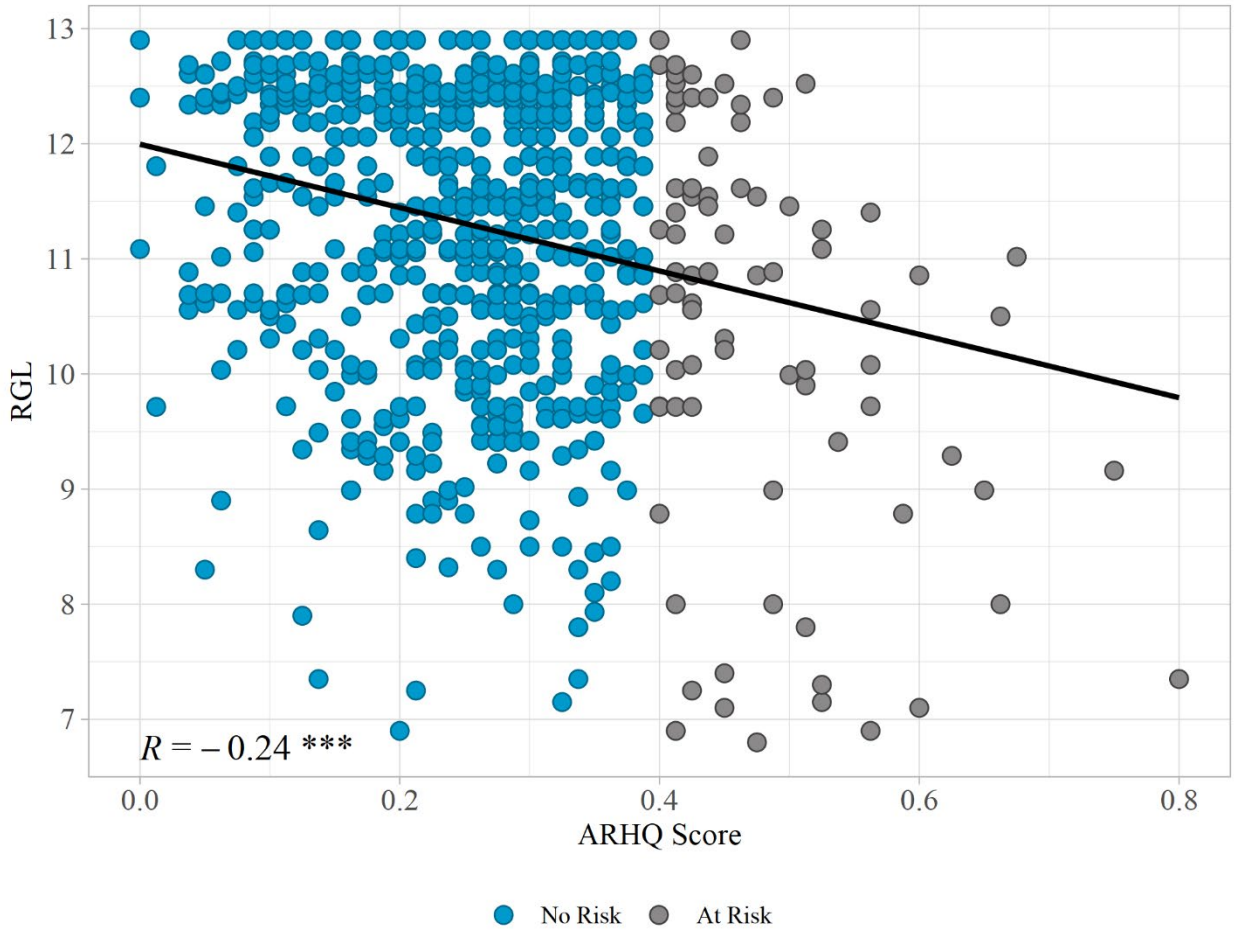
| Score | RGL | Score | RGL |
|---|---|---|---|
| 26 | 6.9 | 56 | 9.4 |
| 27 | 7 | 57 | 9.5 |
| 28 | 7 | 58 | 9.6 |
| 29 | 7.1 | 59 | 9.7 |
| 30 | 7.1 | 60 | 9.8 |
| 31 | 7.2 | 61 | 9.9 |
| 32 | 7.4 | 62 | 10 |
| 33 | 7.4 | 63 | 10.2 |
| 34 | 7.4 | 64 | 10.4 |
| 35 | 7.5 | 65 | 10.5 |
| 36 | 7.6 | 66 | 10.7 |
| 37 | 7.7 | 67 | 10.9 |
| 38 | 7.8 | 68 | 11 |
| 39 | 7.8 | 69 | 11.1 |
| 40 | 7.9 | 70 | 11.3 |
| 41 | 7.9 | 71 | 11.4 |
| 42 | 8 | 72 | 11.6 |
| 43 | 8 | 73 | 11.7 |
| 44 | 8.2 | 74 | 12 |
| 45 | 8.3 | 75 | 12.2 |
| 46 | 8.4 | 76 | 12.4 |
| 47 | 8.5 | 77 | 12.4 |
| 48 | 8.6 | 78 | 12.5 |
| 49 | 8.6 | 79 | 12.6 |
| 50 | 8.7 | 80 | 12.7 |
| 51 | 8.8 | 81 | 12.8 |
| 52 | 8.9 | 82 | 12.9 |
| 53 | 9.1 | 83 | 12.9 |
| 54 | 9.2 | 84 | 12.9 |
| 55 | 9.3 | 85 | 12.9 |

## APPENDIX F – Template

| ALL conditions must be met. | | All thresholds are met | | | | | Copy other thresholds results here. | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | % At Risk | 4.5% | 6.6% | 4.5% | 4.5% | | | | | | |
| **Thresholds for which only ONE must be** | | **One threshold must be met** | | | | | **Copy other thresholds results here.** | | | | | |
| ortho %ile < | 0.02 | ortho %ile < | 0.02 | 0.05 | 0.03 | 0.02 | | | | | | |
| phono %ile < | 0.02 | phono %ile < | 0.02 | 0.05 | 0.03 | 0.02 | | | | | | |
| RGL Equi < | 8.5 | RGL Equi < | 8.5 | 8.0 | 8.0 | 8.0 | | | | | | |
| *Not included in report. | | % At Risk | 6.9% | 10.8% | 7.6% | 6.1% | | | | | | |
| **Thresholds for RGL and either Phono or Ortho** | | **Low RGL and Low Ortho or Low Phono** | | | | | **Copy other thresholds results here.** | | | | | |
| ortho %ile < | 0.2 | ortho %ile < | 0.20 | 0.30 | 0.20 | 0.20 | | | | | | |
| phono %ile < | 0.2 | phono %ile < | 0.20 | 0.30 | 0.20 | 0.20 | | | | | | |
| RGL Equi < | 9.2 | RGL Equi < | 9.2 | 10.0 | 10.00 | 9.5 | | | | | | |
| | | % At Risk | 5.3% | 12.9% | 10.4% | 6.6% | | | | | | |

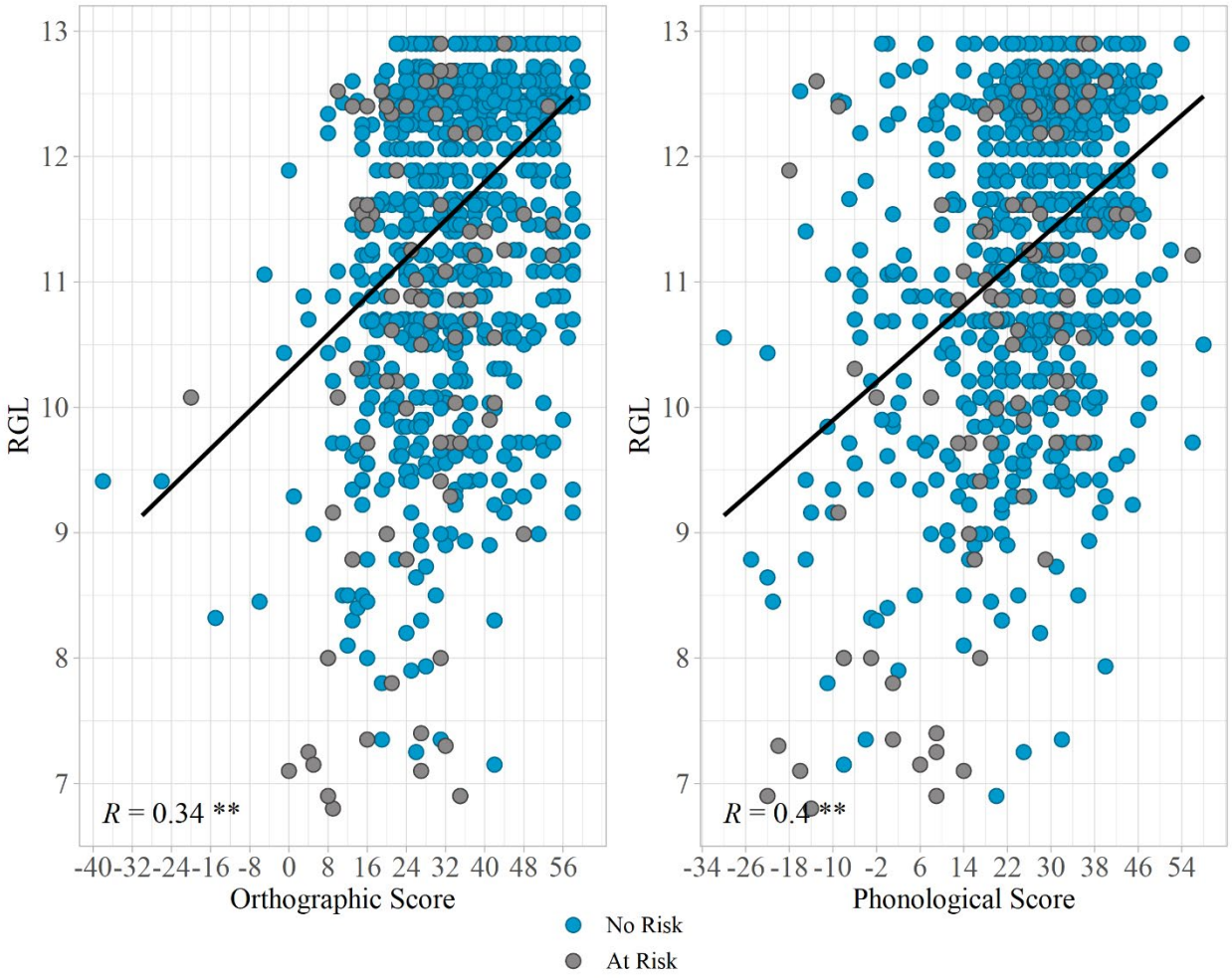
**Figure F -1. Threshold Template**

## APPENDIX G – Additional Graphs




**Correlation significant at the $p$ <.001 level

**Figure G-1. ARHQ and RGL Relationship**




**Correlation significant at the $p$ <.001 level

**Figure G-2. RGL Relationship with Orthographic and Phonological Choice Tests**

## LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

| | |
|---|---|
| AFHRL | Air Force Human Resources Laboratory |
| AFPC/DSYX | Air Force Personnel Center Strategic Research and Assessment Branch |
| AFRAT | Air Force Reading Abilities Test |
| AFS | Air Force Specialty |
| APAT | Applied Performance and Testing |
| ARHQ | Adult Reading History Questionnaire |
| ASHA | American Speech-Language-Hearing Association |
| BMT | Basic Military Training |
| BMTs | Basic Military Trainees |
| HIT | Human Intelligence Task |
| IMC | instructional manipulation check |
| HSD | high school diploma |
| M | mean |
| MTurk | Mechanical Turk |
| N | sample size |
| NCO | Non-Commissioned Officer |
| RGL | Reading Grade Level |
| SD | standard deviation |
| S-RCD | Specific Reading Comprehension Deficit |
| SME | subject matter expert |
| TABE | Test of Adult Basic Education |