# IMPROVING THE FAIRNESS OF COAST GUARD RECRUITMENT & SELECTION WITH THE ASVAB & AFCT:

# A DYNAMIC MEASUREMENT MODELING PARADIGM

Final Report Prepared by Dynamic Measurement, LLC
in Collaboration with the United States Coast Guard

*May 5th, 2022*

Report Authors:

Denis Dumas, PhD
Daniel McNeish, PhD
Yixiao Dong, PhD
Donna Duellberg, EdD

# REPORT DOCUMENTATION PAGE

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



**PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.**

| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| 05052022 | Final | 2007-2021 |

**4. TITLE AND SUBTITLE**
Improving the Fairness of Coast Guard Recruitment & Selection with the ASVAB & AFCT

**5a. CONTRACT NUMBER**
GT&C Number: PR-11639169

**5b. GRANT NUMBER**
NA

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Denis Dumas, PhD
Daniel McNeish, PhD
Yixiao Dong, PhD
Donna Duellberg, EdD

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
OPM/USA Learning Center, Center for Leadership Development

**8. PERFORMING ORGANIZATION REPORT NUMBER**
Order Number: ME21-102

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Office of the Under Secretary of Defense for Personnel and Readiness/M&RA/MPP(AP).
Attn: Dr. Sofiya Velgach, (Assistant Director of Testing Standards, OUSD)

**10. SPONSOR/MONITOR'S ACRONYM(S)**
AP, MPP, M&RA, OUSD (P&R)

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
"A" for public release

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**
A large-scale re-analysis of scores from the Armed Service Vocational Aptitude Battery/Armed Forces Classification Test (ASVAB/AFCT), the primary tools for military recruit accession and training qualification for occupational classification, was conducted using training data from the United States Coast Guard (USCG). The analysis evaluated whether these instruments could be used differently to increase the aperture for different candidate pools, including Underrepresented Minorities (URMs), for qualification into the Coast Guard, as well as qualification for a specific rate qualification. The idea to measure a recruit's learning capacity, by calculating the growth trajectories from in-between test scores, provides the military with more relevant information regarding trainability, which is a new and necessary focus for today's military due to advanced technical training requirements. The research found 10% of Coast Guard recruits re-take the ASVAB/AFCT in order to elevate composite scores, probably with the hopes of qualifying for their occupation of choice. Another 20% of Coast Guard recruits (who are typically within 5 points of qualifying) requested a waiver instead. Underrepresented Minorities (URMs) are more likely to re-test multiple times at a statistically significant level. They are also more likely to see larger improvement gains at a statistically significant level. URM recruits who opted

**15. SUBJECT TERMS**
Learning Capacity; Dynamic Measurement Modeling; Military Psychometrics; Recruitment & Selection; Armed Services Vocational Aptitude Battery; Armed Forces Classification Test

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Dr. Donna Duellberg |
| UNCLAS | UNCLAS | UNCLAS | | 76 | 19b. TELEPHONE NUMBER *(Include area code)* 202-475-5486 |

## Table of Contents

## Executive Summary

Any military force is only as strong as the people who serve within it: individuals who need to be effectively recruited, selected for training schools, and who must pass their training programs successfully. Currently, the Armed Services Vocational Aptitude Battery (ASVAB) and Armed Forces Classification Test (AFCT) are the primary tools for matching recruits to training programs and help ensure that the educational mission of the military operates smoothly. However, amid recruitment shortfalls, as well as on-going conversations about diversity, equity, and inclusion in the military, current ASVAB/AFCT practices may need to be shifted. In this report, we present a large-scale re-analysis of ASVAB/AFCT and training outcome data from The United States Coast Guard (USCG).

In this report, we examine the possibility of shifting ASVAB-based recruitment and selection at the USCG to a dynamic paradigm: a process by which recruits take the test multiple times to demonstrate their learning potential, rather than being matched to training programs after only one test attempt. About 10% of all USCG recruits already elect to take the ASVAB or AFCT multiple times, so we used these data as a natural experiment to examine the possibility of applying a dynamic method to all recruits. We also compared this dynamic method to the current USCG practice of issuing waivers to some recruits if they do not meet the required ASVAB score, and we conducted the entire analysis with an eye toward best-serving women and ethnic/racial minorities, who are historically under-represented in the USCG.

What we found was that women and ethnic/racial minorities were more likely to persist in the recruitment process by taking the ASVAB/AFCT multiple times, and they exhibited statistically significant improvement compared to their White and male peers over the course of their re-takes. Using these multiple ASVAB/AFCT attempts, we were able to apply a cutting-edge statistical process termed Dynamic Measurement Modeling (DMM) to specifically estimate the learning capacity of every USCG recruit. We found that learning capacity scores were fairer—meaning they were less affected by recruits' demographic background—than single timepoint ASVAB scores. In fact, our DMM learning capacity scores improved the fairness of the existing ASVAB scores by nearly 30%. We also found that those recruits who were granted a waiver after only one ASVAB attempt were consistently and substantially more likely to fail their training programs and need to retrain. In comparison, those recruits who earned their way into their training programs by taking the ASVAB multiple times were less likely to need to retrain. This pattern held even after we considered the difficulty of the school each recruit attended, and the effect was particularly pronounced for recruits from historically marginalized groups.

Based on our analysis, we recommend suspending the current USCG practice of waiving ASVAB and AFCT scores after only one attempt, reversing the 10-point reduction of minimum cut-off composite scores, and replacing these with a dynamic perspective on the ASVAB that encourages recruits to retake the test to demonstrate their potential to improve. Our analysis shows that a dynamic paradigm on ASVAB-based recruitment and selection would advance diversity and inclusion within the USCG, improve the rigor with which recruits are selected into training programs, and may be much more cost effective given the high expense with retraining recruits. In this way, dynamic measurement will support the readiness of the US military.

# Final Report Chapter A: Theoretical and Historical Background

## Section A.1: Setting the Stage: Background on Military Psychometrics

The work of the military is inherently psychological—to organize thousands of individuals, each with their own unique life experiences; train them to operate effectively on the tasks required; and then strategically maneuver those thousands of individuals around common goals—necessitates careful attention paid to the psychology of military members and recruits. One way that the military engages in psychologically-oriented efforts is through *psychometrics.* The term psychometrics means 'measuring the mind', and refers to the development, application, interpretation, and statistical analysis of psychological tests (Borsboom, 2005).

The field of psychometrics has deep roots and has been associated with military functioning for millennia. In fact, the first systematically developed and standardized psychometric tests were invented in ancient China over two thousand years ago, and those tests were used to recruit individuals into positions within the government or military (Britannica, 1993). In the United States, the first major psychometric effort to improve military recruitment and training took place over a century ago during World War I (WWI), where academic psychometricians who worked within US universities were tasked with the creation of cognitive batteries to identify the strengths and weaknesses of recruits (i.e., the Alpha and Beta tests; Thorndike, 1919).

Within the emergency context of WWI, severe time constraints were placed on recruits' training and preparation, and therefore the original psychometric tests used by the US military were designed to measure skills and abilities that recruits had *already developed* before their entrance into the military. To put this another way, the initial psychometric batteries of the US military were designed to capture what recruits already knew, so that the military could most effectively utilize the abilities of the individuals who joined the service. During the 20$^{th}$ century, which featured the two largest wars in human history (i.e., WWI and WWII), this psychometric strategy made perfect sense based on the needs of the US military.

However, in the 21$^{st}$ century, the context has changed, as have the needs of the US military. Today, technical and specific training is required for nearly any military recruit in order for them to engage meaningfully with the work that is required of them. So, instead of being concerned with what recruits are already capable of doing, service branches may be more interested in how trainable a recruit may be. In other words, the *learning capacity* of recruits has become potentially more important than their current abilities at the time they are recruited. In addition, it is becoming clear that a focus on already developed abilities is likely to advantage recruits whose life experiences have afforded them the opportunities to develop specific forms of knowledge, while disadvantaging those recruits who come from less privileged backgrounds: whereas a focus on learning capacity would likely allow all recruits to compete on a level playing field (Dumas et al., 2020). However, the principal psychometric recruiting tools used in the US military have not been re-designed or re-analyzed in order to focus on learning capacity. This situation introduces new opportunities to revisit, and enhance, the effectiveness and fairness of military recruitment in the United States.

In this White Paper, we report the findings from a large-scale psychometric research project using recruitment and training outcome data from the US Coast Guard. The overarching intention of this project was to re-formulate and re-analyze these data in order to yield

psychometric quantities that indicate learning capacity, rather than just current time-point ability. This project utilizes a relatively new psychometric technique called *Dynamic Measurement Modeling* (DMM; Dumas & McNeish, 2017) and demonstrates how this technique can be applied to existing data in order to maximize the fairness, validity, and usefulness of psychometric scores.

In the next section (section A.2) of this White Paper, we detail why the particular context of the US Coast Guard was ideal for this research project. Then, we overview the history and purposes of dynamic measurement (section A.3) before explicitly stating the research questions we posed in this study (section A.4). In Chapters B and C of this White Paper, we will provide all Results and Discussion needed to interpret and understand how and why the results of this research are highly meaningful for the US military.

## Section A.2: What is Dynamic Measurement, and How Does it Improve Fairness?

At its core, dynamic measurement refers to psychometric practices that are oriented toward quantifying the learning capacity of individuals, rather than quantifying the level of ability they have already developed (Dumas et al., 2020). What this means more practically is that dynamic measurement procedures require test respondents to take tests multiple times, and for their improvement on those tests to be statistically tracked (McNeish & Dumas, 2017). Then, by extrapolating from their trajectory of improvement on the test, individuals' learning capacity is estimated (Dumas & McNeish, 2017).

### The Development of Dynamic Measurement

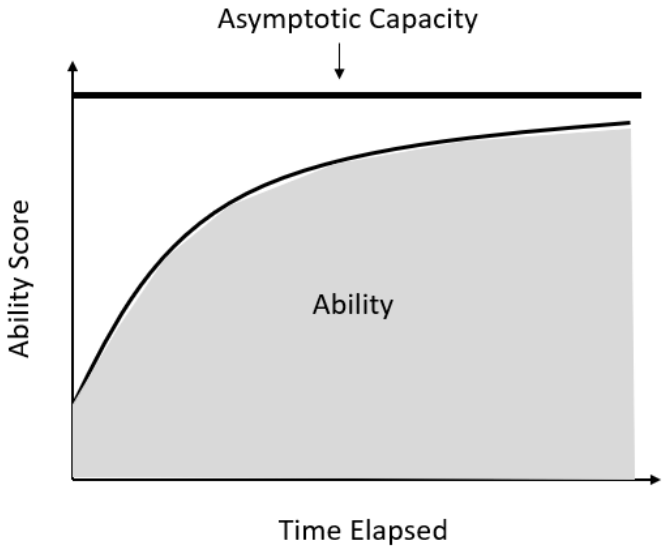
The theoretical concept of dynamic measurement was invented in Israel in the early 1950s by psychologist Reuven Feuerstein (see Feuerstein, 1979). At this time, Feuerstein had the difficult task of sorting child survivors of holocaust concentration camps who had recently immigrated to Israel into grade levels for their schooling. Of course, the terrible deprivation and trauma of the concentration camps meant that these children were essentially never at the grade level that would otherwise be expected based on their age. Interestingly however, when Feuerstein utilized the typical single time-point psychometric tests of his day to determine the grade level of the survivors, he found that he was likely to *under-estimate* their grade-level, and place children in the wrong class. This observation led Feuerstein to a major inference: the child survivors' current knowledge and abilities were severely impacted by the holocaust, but their *learning capacity* was still intact. So, Feuerstein developed the first dynamic tests to begin to quantify children's learning capacity. He then administered the same cognitive batteries to the same children multiple times, and he also taught the children new cognitive strategies in-between the testing occasions. Then, Feuerstein observed the children's trajectory of improvement on the test. Based on the shape and steepness of a child's learning curve, he would infer their learning capacity and sort them into classes that way.

Since Feuerstein's pioneering work, dynamic measurement has been applied many times to address educational psychometric issues, especially regarding students with intellectual disabilities or who suffer from serious traumas (e.g., Haywood & Miller, 2003; Grigorenko, 2009). Application of dynamic measurement has occurred all over the world (especially in Israel and Europe; Elliott, 2003), and researchers have repeatedly documented how attention to students' improvement trajectories, rather than simply their single-timepoint scores, allows for better and more fair uses of psychometric tests. The reason why dynamic measurement is fairer than single-timepoint testing is because the repeated administration of tests allows researcher to track individuals' growth curves. This measurement strategy allows researchers to isolate the capacity of students to learn and grow in response to education, rather than simply observing which students have had more opportunity for learning in their past.

Importantly, it has been known among psychologists for many decades that individuals tend to improve and learn in a nonlinear way: they initially grow rapidly in response to the instruction, but their learning gains decelerate such that their learning curve eventually levels out toward an upper asymptote (Dumas et al., 2020). One particularly promising aspect of dynamic

measurement is that, if respondents take a test enough times, the upper asymptote on their learning curve—or their learning capacity for that particular test—can be predicted and quantified. Please see Figure 1 below for a visual representation of how nonlinear learning curves allow for learning capacity to be observed. The space below the nonlinear growth curve is realized ability and the asymptotic line at the top is the capacity.

Figure 1.
*Theoretical Depiction of Components of Potential*




**Dynamic Measurement Modeling in our Laboratory**

Our research group has pioneered statistical techniques (i.e., DMM) for tracking individual test respondent's learning trajectories and estimating their learning capacities. We have built a body of evidence that strongly suggests that learning capacity scores are much fairer representations of individuals' capabilities than are single-timepoint test scores (Dumas & McNeish, 2017). Our work has incorporated students at many age-levels ranging from preschool (Dumas, McNeish, Sarama, et al., 2019; Dong et al., 2022), elementary and middle school (McNeish & Dumas, 2020; McNeish, Harring, & Dumas, under review), through medical school (Dumas, McNeish, Schreiber-Gregory, et al., 2019; McNeish, Dumas, Torre, & Rice, 2022).

In all these cases, our modern statistical work confirmed Feuerstein's theoretical intuition: dynamic measurement is a much fairer method of quantifying student capabilities than single-timepoint testing. In another major paper that won the Tanaka award for best paper in 2021 from the Society for Multivariate Experimental Psychology (i.e., McNeish et al., 2021), we were able to show that our DMM statistical method was able to extrapolate the level of cognitive ability an individual would exhibit at age 72 based only on their learning trajectories during their childhood and teenage years. For these reasons, we strongly believe that DMM is likely to be an effective method for improving the fairness of testing programs in a variety of areas: by shifting the emphasis from current abilities and knowledge to individual's capacity to improve, we can

elucidate the learning potential of all people, even those that have received somewhat limited learning opportunities over the course of their life.

**DMM as a Route Toward Equity and Inclusion**

Importantly, in the US, opportunities to learn cognitive skills have historically been unequally distributed across the population (Flores, 2007). For example, US citizens with ethnic/racial heritage in Europe (i.e., those who are White) have tended to receive more and better opportunities to learn on average than other US citizens who come from other ethnic/racial groups (e.g., African American; Kuhfeld et al., 2018). However, as many people intrinsically understand, and our research has empirically shown (Dumas & McNeish, 2017), the learning capacity of these groups is essentially equal. For this reason, we strongly advocate for a shift in US psychometric practice away from single-timepoint testing and toward a dynamic approach to testing. Based on our data, we believe that a much fairer and more valid psychometric practice could emerge and become typical in the US. The use of dynamic measurement is therefore profoundly motivated by a need for fairness in psychometric testing programs. The purpose of this current research project is to determine whether this conjecture is true in the context of US military recruitment psychometrics, and specifically within the context of recruitment and selection for the US Coast Guard.

## Section A.3: Current Issues and Opportunities in Recruitment and Selection for the United States Coast Guard

The Armed Services Vocation Aptitude Battery (ASVAB) consists of multiple sub-tests in nine areas: general science, auto and shop information, mechanical comprehension, assembling objects, electronics information, word knowledge, paragraph comprehension, arithmetic reasoning, and mechanical comprehension. The scores help the armed services in their selection and classification of personnel (https://www.offcialasvab.com). The scores from four sub-tests (word knowledge, paragraph comprehension, arithmetic reasoning, and mathematics knowledge) are combined to the produce the "Armed Forces Qualification Test" or AFQT score. Congress mandates the use of the AFQT to access overall "trainability" and entrance eligibility into the armed services (Bayroff & Fuchs, 1970). Requirements of Title 10, United States Code, further use the AFQT to establish adherence to qualitative distribution benchmarks for AFQT categories and education credential tiers. For example, the total number of Category IV enlistments (10-30 AFQT percentile) may not exceed 20% of the total number of active duty accessions for each Armed Service, and 60% of those accessions must be categorized as AFQT IIIA or higher, with percentile scores of at least 50 (DoDI 1145.01, 2013). Based on this mandate, each of the armed services sets its own AFQT minimum. In accordance with Commandant Instruction M1100.2G, dated August 2021, the Coast Guard's Recruiting Manual, an AFQT score of "36" is the minimum score the Coast Guard used for high school graduate recruits, but effective January 2022, the AFQT was reduced to "32."

There are 19 Coast Guard occupation specialties for its active duty, and each of these occupational specialties possesses its own ASVAB or Armed Forces Classification Test (AFCT) minimum composite score (Table 1). The Coast Guard Reserve has three additional occupations (Table 2). Both tables reflect minimum cut-off composites effective November 2021 after the most recent policy change, which cut minimum composite scores by 10 points across rates (USCG, 2021). Table 3 depicts minimum composite scores for "A" school prior to the policy change. The ASVAB and AFCT both assess aptitude for occupational success. The difference between the two is that while versions of the ASVAB are taken by civilians, the AFCT is taken by individuals who are already enlisted in the service. The three versions of the ASVAB target a different population: the ASVAB Career Explorer Program (CEP) is predominately taken by high school or college students, who use the tool to "explore" careers; the Computer Adaptive Test (CAT) version, is taken at Military Entrance Processing Stations (MEPS) and is solely used for enlistment; and the Mobile Examination Test (MET)-site ASVAB is a pencil and paper version taken at decentralized locations usually placed throughout cities. Regardless of which ASVAB a recruit takes, in addition to meeting the armed services' qualifying AFQT score, he or she would also need to meet other minimum composite scores to qualify for training in a specific occupation.

The Coast Guard is unique as most of its recruits enter the service as "non-rates" and stay in this "apprentice" status until they attend an "A" school or other formal on-the-job training program (e.g., Striker), and become rate qualified. The "non-rate" window provides an opportunity for Coast Guard recruits to observe many occupations before selecting their military life work, but that window eventually closes. If a non-rate has selected a rate and is "ready" to attend training, he or she must first meet the minimum composite score. When the minimum composite score is

not met, the individual can retest to increase the score or request a waiver. Under the newest policy change announced in ACN 112/21 the Coast Guard message system, Commanding Officers (COs) or Officers in Charge (OICs) can waive the minimum composite scores up to 10 points, and Rating Force Master Chiefs (RFMCs) can grant waivers for 11 or more points. With this policy change, service members who missed meeting the minimum composite scores by up to 20 points prior to November 2021, can now meet the minimum "A" school requirement. Consequently, the expectation for AFCT test/retesting should greatly decrease.

Table 1.
*Rate ASVAB/AFCT Requirements for Coast Guard Active Duty, effective 01 November 2021*

| **RATE** | **ASVAB/AFCT requirement** | **RATE** | **ASVAB/AFCT requirement** |
|---|---|---|---|
| Avionics Electrical Technician (AET) | AFQT = 65 OR MK+EI+GS = 162 w/ AR minimum of 52 | Aviation Maintenance Technician (AMT) | AFQT = 65 OR AR+MC+AS+EI = 210 w/ AR minimum of 52 |
| Aviation Survival Technician (AST) | AFQT = 65 OR VE+MC+AS = 152 w/ AR minimum of 52 | Boatswain's Mate (BM) | AR+VE = 90 |
| Culinary Specialist (CS) | VE+AR = 95 | Damage Controlman (DC) | VE+MC+AS = 145 |
| Electrician's Mate (EM) | MK+EI+GS = 143 w/ AR minimum of 52 | Electronics Technician (ET) | AFQT = 65 OR MK+EI+GS = 162 w/ AR minimum of 52 |
| Gunner's Mate (GM) | AR+MK+EI+GS = 199 | Health Service Technician (HS) | VE+MK+GS+AR=197 w/ AR minimum of 50 |
| Intelligence Specialist (IS) | AR+VE = 99 | Information System Technician (IT) | AFQT = 65 OR MK+EI+GS = 162 |
| Maritime Enforcement Specialist (ME) | AR+VE = 90 | Machinery Technician (MK) | AR+MC+AS = 144 OR VE+AR = 95 |
| Marine Science Technician (MST) | VE+AR = 104 w/ MK minimum of 56 | Operations Specialist (OS) | VE+AR = 95 |
| Public Affairs Specialist (PA) | VE+AR = 99 w/ VE minimum of 54 | Storekeeper (SK) | VE+AR = 95 w/ VE minimum of 51 |
| Yeoman (YN) | VE+AR = 95 | | |

Table 2.
*Rate ASVAB/AFCT Requirements for Coast Guard Reservist*

| **RATE** | **ASVAB/AFCT requirement** | **RATE** | **ASVAB/AFCT requirement** |
|---|---|---|---|
| Diver (DV) | AR+WK = 104 AND MC = 50 | Investigator (IV) & Musician (MUS) | NA |
| | | | |

Table 3.
*Rate ASVAB/AFCT Requirements for Coast Guard Active Duty, prior 01 November 2021*

| | SCORE | MIN | | QUAL? | | SCORE | MIN | | QUAL? |
|---|---|---|---|---|---|---|---|---|---|
| **AET** | | | | | **HS** | | | | |
| MK+EI+GS= 172 | 0 | AR of 52 | 0 | NO | VE+MK+GS+AR = 207 | 0 | AR of 50 | 0 | NO |
| or AFQT of 65 or above | 0 | | | | **IS** | | | | |
| **AMT** | | | | | VE+AR= 109 | 0 | | | NO |
| AR+MC+AS+EI= 220 | 0 | AR of 52 | 0 | NO | **IT** | | | | |
| or AFQT of 65 or above | 0 | | | | MK+EI+GS= 172 | 0 | AR of 52 | 0 | NO |
| **AST** | | | | | or AFQT of 65 or above | 0 | | | |
| VE+MC+AS= 162 | 0 | AR of 52 | 0 | NO | **ME** | | | | |
| or AFQT of 65 or above | 0 | | | | VE+AR= 100 | 0 | | | NO |
| **BM** | | | | | **MK** | | | | |
| VE+AR= 100 | 0 | | | NO | AR+MC+AS= 154 | 0 | | | NO |
| **DC** | | | | | or VE+AR = 105 | 0 | | | |
| VE+MC+AS= 155 | 0 | | | NO | **MST** | | | | |
| **EM** | | | | | VE+AR= 114 | 0 | MK of 56 | 0 | NO |
| MK+EI+GS= 153 | 0 | AR of 52 | 0 | NO | **OS** | | | | |
| **ET** | | | | | VE+AR= 105 | 0 | | | NO |
| MK+EI+GS= 172 | 0 | AR of 52 | 0 | NO | **PA** | | | | |
| or AFQT of 65 or above | 0 | | | | VE+AR= 109 | 0 | VE of 54 | 0 | NO |
| **CS** | | | | | **SK** | | | | |
| VE+AR= 105 | 0 | | | NO | VE+AR= 105 | 0 | VE of 51 | 0 | NO |
| **GM** | | | | | **YN** | | | | |
| AR+MK+EI+GS = 209 | 0 | | | NO | VE+AR= 105 | 0 | | | NO |

The Nov 2021 policy change and Dec 2021 CG-1 Directive to decrease the AFCT to “32” from “36” were implemented to open the aperture for “A” school qualification as well as qualification into the Coast Guard. This was in response to alleviating the Coast Guard’s projected difficulty in meeting mission requirements, as well as these findings: there is a disproportionate number of minority groups requesting waivers (Table 4); the rating qualification percentages without any waiver are lower for minority groups (Table 5); and in terms of how this data translates into opportunity, the number of rating choices for each individual population, if no waivers are in effect, are less for minority groups (Table 6).

Table 4.
*Percentage of Waiver Requests in Minority Groups (Lord & Mayer, 2020, Reproduction Permission Received)*

| Group | FY14-FY19 Accessions | FY18-FY20 Waivers Requested |
|---|---|---|
| White Non-Hispanic Men | 55.5% | 29.6% |
| Women | 15.9% | 24.3% |
| African American / Black | 7.25% | 17.1% |
| Hispanic | 19.6% | 32.2% |

Table 5.
*Rating Qualification with No Waiver (Lord & Mayer, 2020, Reproduction Permission Received)*

| Red cells less than 80% of max group. Per: 29 CFR 1607.4 | Overall | AET | AMT | AST | BM | DC | EM | ET | CS | GM | HS | IS | IT | MK | ME | MST | OS | PA | SK | YN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 18431 | 6533 | 7772 | 8303 | 15738 | 12201 | 11545 | 6533 | 12609 | 13140 | 13008 | 10099 | 6533 | 14635 | 15738 | 5716 | 12609 | 8762 | 11749 | 12609 |
| | | 35% | 42% | 45% | 85% | 66% | 63% | 35% | 68% | 71% | 71% | 55% | 35% | 79% | 85% | 31% | 68% | 48% | 64% | 68% |
| Gender | | | | | | | | | | | | | | | | | | | | |
| F | 2936 | 22% | 19% | 25% | 81% | 39% | 50% | 22% | 61% | 56% | 62% | 47% | 22% | 64% | 81% | 26% | 61% | 41% | 57% | 61% |
| M | 15495 | 38% | 47% | 49% | 86% | 71% | 65% | 38% | 70% | 74% | 72% | 56% | 38% | 82% | 86% | 32% | 70% | 49% | 65% | 70% |
| Race | | | | | | | | | | | | | | | | | | | | |
| American Indian or Alaska Native | 411 | 33% | 47% | 50% | 90% | 73% | 69% | 33% | 73% | 76% | 73% | 58% | 33% | 86% | 90% | 30% | 73% | 49% | 69% | 73% |
| Asian | 439 | 44% | 46% | 45% | 89% | 68% | 72% | 44% | 76% | 78% | 80% | 66% | 44% | 83% | 89% | 44% | 76% | 55% | 69% | 76% |
| Black or African American | 1336 | 15% | 14% | 15% | 71% | 32% | 37% | 15% | 46% | 45% | 48% | 32% | 15% | 54% | 71% | 15% | 46% | 27% | 42% | 46% |
| Native Hawaiian or Other Pacific Islanc | 315 | 24% | 29% | 30% | 77% | 51% | 50% | 24% | 58% | 59% | 60% | 42% | 24% | 69% | 77% | 20% | 58% | 33% | 50% | 58% |
| Other/Multi-Race | 1312 | 25% | 29% | 32% | 79% | 51% | 53% | 25% | 59% | 62% | 61% | 46% | 25% | 69% | 79% | 25% | 59% | 40% | 55% | 59% |
| White | 14618 | 38% | 46% | 49% | 87% | 71% | 66% | 38% | 71% | 74% | 73% | 58% | 38% | 83% | 87% | 33% | 71% | 50% | 67% | 71% |
| Hispanic | | | | | | | | | | | | | | | | | | | | |
| N | 14815 | 38% | 46% | 49% | 87% | 70% | 66% | 38% | 71% | 74% | 73% | 58% | 38% | 82% | 87% | 33% | 71% | 51% | 67% | 71% |
| Y | 3616 | 23% | 26% | 28% | 79% | 50% | 51% | 23% | 57% | 59% | 60% | 42% | 23% | 68% | 79% | 21% | 57% | 35% | 51% | 57% |

Table 6.
*Amount of Ratings with No Waiver (Lord & Mayer, 2020, Reproduction Permission Received)*

| | Amount of Ratings | | | | |
|---|---|---|---|---|---|
| No Waiver | 0 | 1-5 | 6-10 | 11-18 | 19 |
| Overall | 10% | 18% | 13% | 39% | 21% |
| Gender | | | | | |
| F | 17% | 21% | 17% | 34% | 12% |
| M | 8% | 18% | 13% | 39% | 22% |
| Race | | | | | |
| American Indian or Alaska Native | 5% | 17% | 15% | 43% | 19% |
| Asian | 8% | 12% | 13% | 39% | 28% |
| Black or African American | 23% | 29% | 16% | 26% | 6% |
| Native Hawaiian or Other Pacific Islander | 16% | 23% | 18% | 31% | 12% |
| Other/Multi-Race | 15% | 23% | 14% | 34% | 13% |
| White | 8% | 17% | 13% | 40% | 23% |
| Hispanic | | | | | |
| N | 8% | 17% | 12% | 40% | 23% |
| Y | 15% | 24% | 17% | 32% | 12% |

Decreasing minimum composite cut-offs by 10 points across the Coast Guard OR if a 10-point waiver is in effect, would change the percentages favorably (Table 7) and qualification for all USCG rates increases across the board (Table 8):

Table 7.
*Amount of Ratings with 10 Point Waiver (Lord & Mayer, 2020, Reproduction Permission Received)*

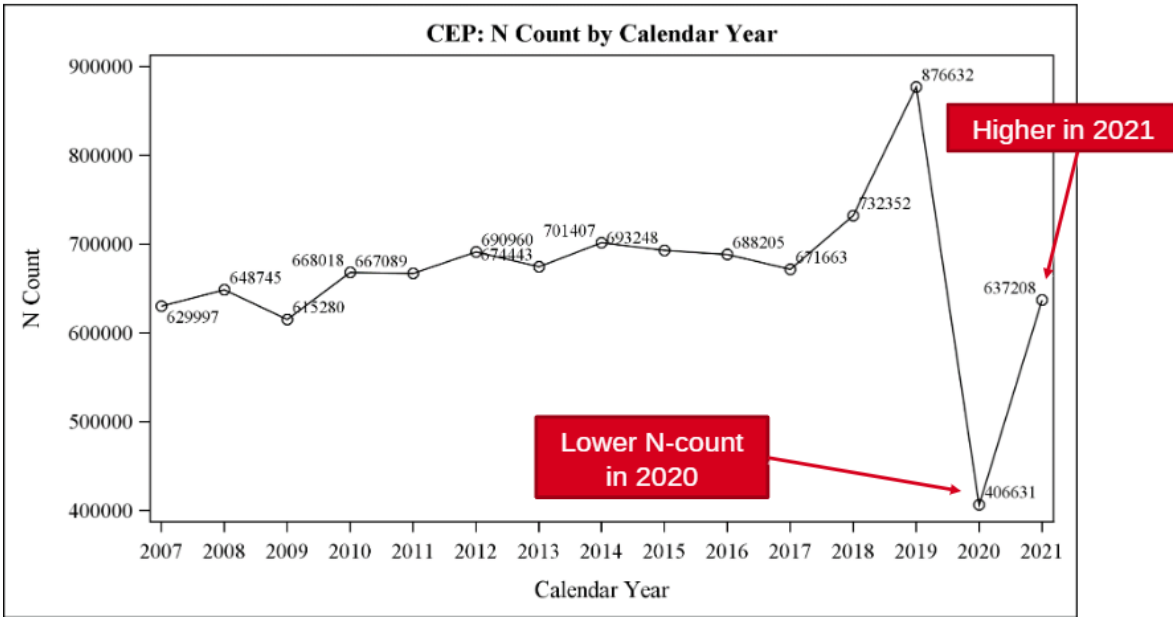
| | Amount of Ratings | | | | |
|---|---|---|---|---|---|
| 10 Points with Subtests | 0 | 1-5 | 6-10 | 11-18 | 19 |
| Overall | 0% | 3% | 11% | 42% | 44% |
| Gender | | | | | |
| F | 1% | 4% | 18% | 50% | 27% |
| M | 0% | 2% | 10% | 40% | 48% |
| Race | | | | | |
| American Indian or Alaska Native | 0% | 1% | 7% | 47% | 45% |
| Asian | 1% | 3% | 7% | 42% | 48% |
| Black or African American | 1% | 7% | 26% | 48% | 18% |
| Native Hawaiian or Other Pacific Islander | 1% | 4% | 15% | 52% | 28% |
| Other/Multi-Race | 1% | 5% | 16% | 46% | 32% |
| White | 0% | 2% | 9% | 40% | 48% |
| Hispanic | | | | | |
| N | 0% | 2% | 9% | 40% | 48% |
| Y | 1% | 5% | 18% | 48% | 29% |

Table 8.
*Increases in Qualification Percentages between 10 Point Waiver and No Waiver (Lord & Mayer, 2020, Reproduction Permission Received)*

| | Increase in Qualificaton Percentage between 10 point and No waiver | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | AET | AMT | AST | BM | DC | EM | ET | CS | GM | HS | IS | IT | MK | ME | MST | OS | PA | SK | YN |
| | 18431 | 4008 | 3225 | 4388 | 2614 | 3358 | 4952 | 4008 | 5151 | 2711 | 3441 | 6148 | 4008 | 3343 | 2614 | 6112 | 5151 | 5541 | 5037 | 5151 |
| | | 22% | 17% | 24% | 14% | 18% | 27% | 22% | 28% | 15% | 19% | 33% | 22% | 18% | 14% | 33% | 28% | 30% | 27% | 28% |
| Gender | | | | | | | | | | | | | | | | | | | | |
| F | 2936 | 20% | 13% | 19% | 18% | 27% | 32% | 20% | 34% | 20% | 24% | 37% | 20% | 31% | 18% | 32% | 34% | 32% | 32% | 34% |
| M | 15495 | 22% | 18% | 25% | 13% | 17% | 26% | 22% | 27% | 14% | 18% | 33% | 22% | 16% | 13% | 33% | 27% | 30% | 26% | 27% |
| Race | | | | | | | | | | | | | | | | | | | | |
| American Indian or Alaska Native | 411 | 25% | 20% | 27% | 10% | 18% | 27% | 25% | 25% | 17% | 20% | 35% | 25% | 13% | 10% | 36% | 25% | 33% | 25% | 25% |
| Asian | 439 | 22% | 13% | 22% | 11% | 18% | 19% | 22% | 20% | 12% | 12% | 25% | 22% | 14% | 11% | 31% | 20% | 25% | 22% | 20% |
| Black or African American | 1336 | 16% | 13% | 21% | 27% | 25% | 36% | 16% | 46% | 21% | 27% | 44% | 16% | 39% | 27% | 27% | 46% | 33% | 41% | 46% |
| Native Hawaiian or Other Pacific Islanc | 315 | 18% | 15% | 24% | 22% | 30% | 33% | 18% | 35% | 20% | 22% | 39% | 18% | 27% | 22% | 36% | 35% | 35% | 35% | 35% |
| Other/Multi-Race | 1312 | 20% | 16% | 23% | 20% | 24% | 30% | 20% | 35% | 16% | 21% | 37% | 20% | 26% | 20% | 30% | 35% | 31% | 32% | 35% |
| White | 14618 | 22% | 18% | 24% | 13% | 17% | 26% | 22% | 26% | 14% | 18% | 32% | 22% | 16% | 13% | 34% | 26% | 30% | 26% | 26% |
| Hispanic | | | | | | | | | | | | | | | | | | | | |
| N | 14815 | 22% | 18% | 24% | 13% | 17% | 26% | 22% | 26% | 14% | 18% | 32% | 22% | 16% | 13% | 34% | 26% | 29% | 26% | 26% |
| Y | 3616 | 20% | 16% | 25% | 21% | 25% | 31% | 20% | 37% | 18% | 23% | 40% | 20% | 27% | 21% | 32% | 37% | 33% | 34% | 37% |

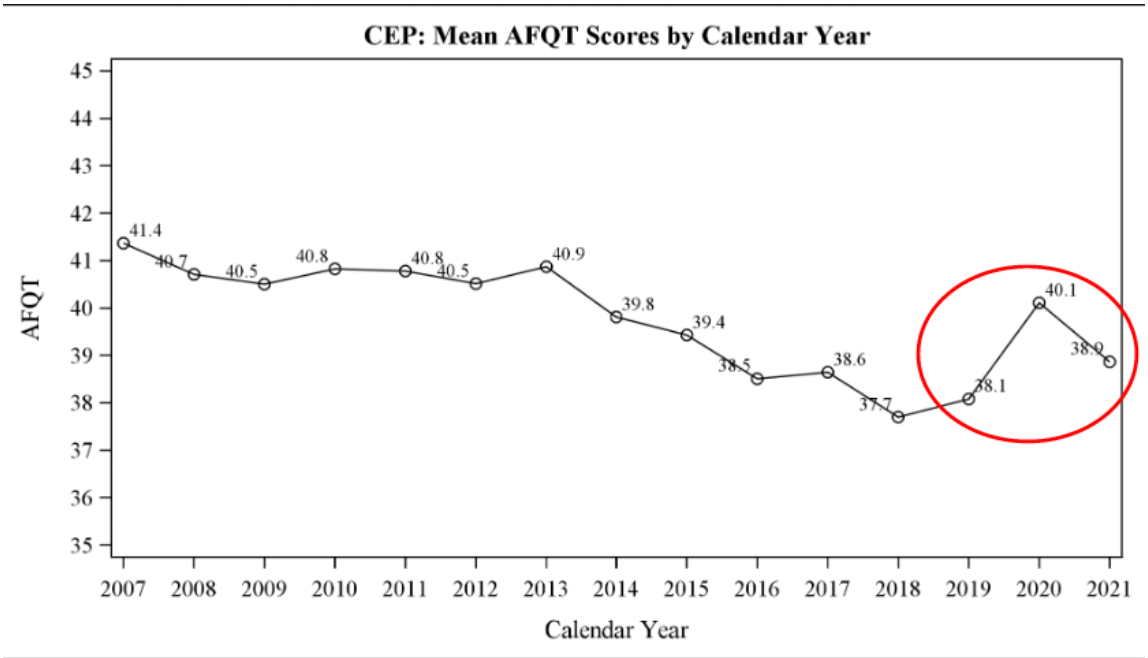
All services struggle with selection and classification, albeit for different reasons and at different times. For example, when the economy is good and unemployment is low, the armed services typically experience recruitment shortages (Stafford & Griffis, 2008). Most millennials have no interest in serving in the military (Colford & Sugerman, 2016), and the "graying of America" presents a significant obstacle (Quester, 2005). Most recently, the coronavirus disease impacted the services, too, with most experiencing far fewer accessions (Calkins & Asch, 2022). This is partly because the number of high school students taking the ASVAB CEP dropped to almost half in 2020 (see Figure 2).

Figure 2.
*CEP: Number Count by Calendar Year (Yin, 2022, Reproduction Permission Received)*



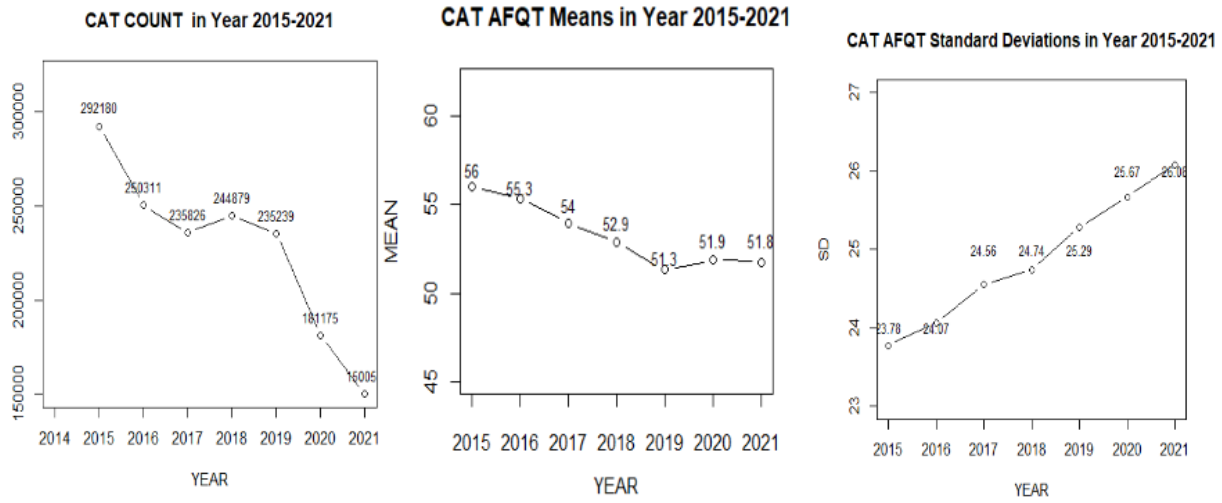

On the upside, the students who took the test in 2020 did better on the AFQT (see Figure 3).

Figure 3.
*CEP: Mean AFQT Scores by Calendar year (Yin, 2022, Reproduction Permission Received)*



The ASVAB CAT saw a similar downward trend in the number of test takers, but whereas high school students increased the mean AFQT, potential recruits in 2020 did only slightly better than 2019 test takers (see Figure 4).

Figure 4.
*CAT-ASVAB ETP Case Count & AFQT Score Means and Standard Deviations, by Year: 2015–2021 (Yao, 2022, Reproduction Permission Received)*




It is a complex task for the military services to not only meet recruiting numbers, but simultaneously improve representation of women and racial and ethnic minorities within their ranks. The Coast Guard, along with the other armed services, has spent considerable time reviewing eligibility requirements and assessing if these are serving as barriers for under-represented minorities (Lim et al, 2021). In 2021, the Coast Guard implemented changes to accept lower ASVAB/AFCT scores: specifically, waivers were increased while it simultaneously lowered the minimum qualification composite scores by 10 points across the board for entry into "A" schools for occupational training (Connell, 2021).

The services struggle with different ideas on how to expand eligibility into the services and classification into occupations. Many of the most popular proposed solutions revolve around changes to the administration of the ASVAB or AFCT to improve scores including, authorize calculators, transcribe into Spanish and other languages, combine sub-tests differently to find alternate composite scores, or expand waivers so new minimums could be used. However, most of these solutions are continuously rejected because they introduce no real change (e.g., adding a calculator, if proved to be an advantage, would advantage all test-takers and after the test was re-normed, the distribution of where one would fall would likely be the same); or they can't be done economically (e.g., although multilingual students perform better in their common language than the dominant language of instruction [Canz et al., 2021], creating and maintaining the ASVAB/AFCT in a different language would be very expensive). Plus, while the "Every Student Succeeds Act" mandates assessments in students' first languages, no such law extends to military recruitment. Finally, it is difficult to guarantee that even a well-translated test can produce a validly comparative score (Hambleton, 2002).

What remains to be seen is if increasing waivers increases attrition and lowers graduation rates. Clearly, there is risk in increasing waivers. If the minimums are really minimums required to qualify for not just entry into an "A" school, but also to have a high probability to graduate from the school, the resulting data might make the lack of diversity and inclusion ultimately worse, not better (e.g., increasing waivers may decrease graduation rates, which could increase attrition). Unless training over time has gotten easier, this new adoption of increased waivers implicates years of discriminatory practices since the pre-existing minimums were never really minimums but inflated minimums: such a practice could have led to years of disparate impact for minorities (think Griggs v. Duke Power Co).

This current dynamic measurement research offers the Coast Guard (as well as the other services) an expanded use of the ASVAB to estimate learning capacity and use that score for determining military enlistment eligibility and occupation classification eligibility. Lowering standards perpetuates the stereotype that minorities need charity, instead of an opportunity. Adopting larger waivers or lower standards on the existing ASVAB scores seems counter-intuitive because it could harbor sentiments of tokenism among members of underrepresented groups or could trigger resentment towards underrepresented groups if they are viewed as receiving handouts because of the demographic characteristics.

Because it is vital for the Coast Guard to close its workforce gaps, mitigate future personnel shortages, and foster diversity and inclusion, a real solution is necessary. This research offers that; it presents a different solution for the USCG based on data driven analyses without compromising standards and minimizes risk.

## Section A.4: Specific Goals for Current Project

Given the theoretical and historical background of dynamic measurement that has been reviewed here, as well as the specific context and recruiting and selection needs of the US Coast Guard, this current research project posited several specific empirical goals.

### Goal One: Data Organization

At the outset of this project, recruitment data for the US Coast Guard was not organized in a way that allowed for a full comprehensive dynamic measurement analysis. Therefore, this project required detailed work related to data organization and merging that produced a newly formulated dataset for analysis. Details about how this dataset was constructed and how to interpret each of the variables within that dataset are found in section B.1 of this report.

### Goal Two: Dynamic Measurement Modeling

Every psychometric dataset presents unique challenges and opportunities, and therefore DMM methodology must necessarily be tailored to each project our lab undertakes. In this project, we formulated a DMM to meet the specific needs of the US Coast Guard using statistical techniques that had never been applied in this context before. Details of this modeling phase of the project are present in section B.2 of this report.

### Goal Three: Interpreting Fairness, Validity, and Consequences

After developing and fitting the novel DMM model in this project, our team was able to closely interpret the coefficients from the model and analyze how much the DMM improved the fairness of the recruitment testing program over and above a single-timepoint testing paradigm. We also were able to investigate the ways in which DMM parameters predicted training outcomes for Coast Guard recruits to determine how DMM coefficients can be useful for recruiting purposes. Results of these goals are present in sections B.3 and B.4 of this report. To explore how DMM compares to proposals to increase waivers, Section B.5 reports on differences in training outcomes for recruits who received a waiver after one ASVAB attempt and recruits with multiple ASVAB attempts who never received a waiver.

### Goal Four: Formulating Recommendations for Future Work

The current project represents the first-ever application of DMM to military recruitment psychometrics in the US. For that reason, a major goal of the project was to determine whether and how it was feasible or worthwhile to continue to apply DMM in this context. In Chapter C of this White paper (which has four subsections to its Discussion), we carefully consider this important question and posit our team's recommendations for future dynamic measurement work within the US military.

# Final Report Chapter B: Methodology and Results

## Section B.1: Data Context and Descriptive Patterns

The present project featured an empirical analysis of data from the United States Coast Guard (CG) data which contains the Armed Services Vocational Aptitude Battery (ASVAB)/Armed Forces Classification Test (AFCT) scores of each recruit, demographics (i.e., race, gender, and Hispanic or not), and other training relevant information (e.g., ID records, names of training schools, and course start & end dates).

### Analytic Sample

The original dataset contains 21,027 USCG recruits who entered training schools from June 2013 to May 2021. However, a portion of these participants came with missing data, anomalous ASVAB scores (e.g., 0), incorrect test-taking dates (e.g., 01-Jan-1951), and/or other entry errors. For the later DMM modeling and fairness analyses in Sections B.2 through B.5, we have cleaned and re-organized the data received from the Coast Guard. All the cleaned data are stored in the *Master Data & Dictionary .xlsx* file, along with a data dictionary created by us.

The final analytic sample featured a total of 18,210 recruits. Table 9 shows the self-reported demographic breakdown of the sample. As can be seen, the sample included 16.2% female and 17.5% Hispanic/Latinx recruits. The majority (73.9%) of participants reported their race as White; 6% reported their race as African American; 2.2% reported their race as Asian; 1.9% reported their race as American Indian/Alaska Native; 1.5% of participants reported their race as Native Hawaiian/Pacific Islander; 2.9% of them were multi-race, and 11.6% had no records of race.

Table 9.
*Demographics of Sample*

| Variables | Group | N | Percentage |
|---|---|---|---|
| *Gender* | | | |
| | Male | 15262 | 83.8 |
| | Female | 2948 | 16.2 |
| *Race* | | | |
| | White | 13462 | 73.9 |
| | Black or African American | 1094 | 6 |
| | Asian | 398 | 2.2 |
| | American Indian/Alaska Native | 353 | 1.9 |
| | Native Hawaiian/Pacific Islander | 267 | 1.5 |
| | Multi-Race | 526 | 2.9 |
| | No records (missing data) | 2110 | 11.6 |
| *Hispanic or Latinx* | | | |
| | No | 12607 | 69.2 |
| | Yes | 3191 | 17.5 |
| | No records (missing data) | 2412 | 13.2 |
| Total | | 18210 | 100 |

## ASVAB Test Battery

To estimate learning capacity of each USCG recruit via DMM, we utilized the ASVAB measures (Jensen, 1985). In overview, the ASVAB test has demonstrated sound psychometric and statistical characteristics (e.g., good reliability; U. S. Department of Defense, 1999). Despite previous studies supporting some aspects of validity with the ASVAB (see a review of ASVAB validity, Welsh et al., 1990), the consequential validity (or test fairness) of the ASVAB remains a concern.

As displayed in Table 10, the ASVAB recruits receive nine sub-scores and an Armed Forces Qualification Test (AFQT) score. The AFQT is a composite score (in percentile rank) of Arithmetic Reasoning, Mathematics Knowledge, Paragraph Comprehension, and Word Knowledge. Verbal Expression is a composite score of Word Knowledge and Paragraph Comprehension. Given that the AFQT score has been used as an essential qualification to determine whether an recruit may enter various military services (Kapp, 2002), the current project uses the AFQT score to estimate participants' learning capacity. Notably, the ASVAB measure or score in the later sections of this report (B.2 to B.5 and Chapter C) generally refers to the AFQT composite score.

Table 10.
*ASVAB Subtests and Armed Forces Qualification Test Components*

| **Individual ASVAB Scores** | **AFQT** |
|---|---|
| Arithmetic Reasoning (AR) | X |
| Automotive and Shop Information (AS) | |
| Assembling Objects (AO) | |
| Electronics Information (EI) | |
| General Science (GS) | |
| Mechanical Comprehension (MC) | |
| Mathematics Knowledge (MK) | X |
| Paragraph Comprehension (PC) | X |
| Verbal Expression (VE) = (WK)+(PC) | |
| Word Knowledge (WK) | X |

## Outcome Measures

We performed multiple analyses to examine the validity of the capacity scores as well as the incremental value of using DMM scores to predict recruits' training outcomes (presented in Sections B.3 through B.5 below). To achieve this goal, we have generated distal outcome variables with the USCG data.

### *Recycling*

Recycling status indicates whether a person needs to remain in their training longer than originally planned. Given that each training school has a typical training length (e.g., Aviation

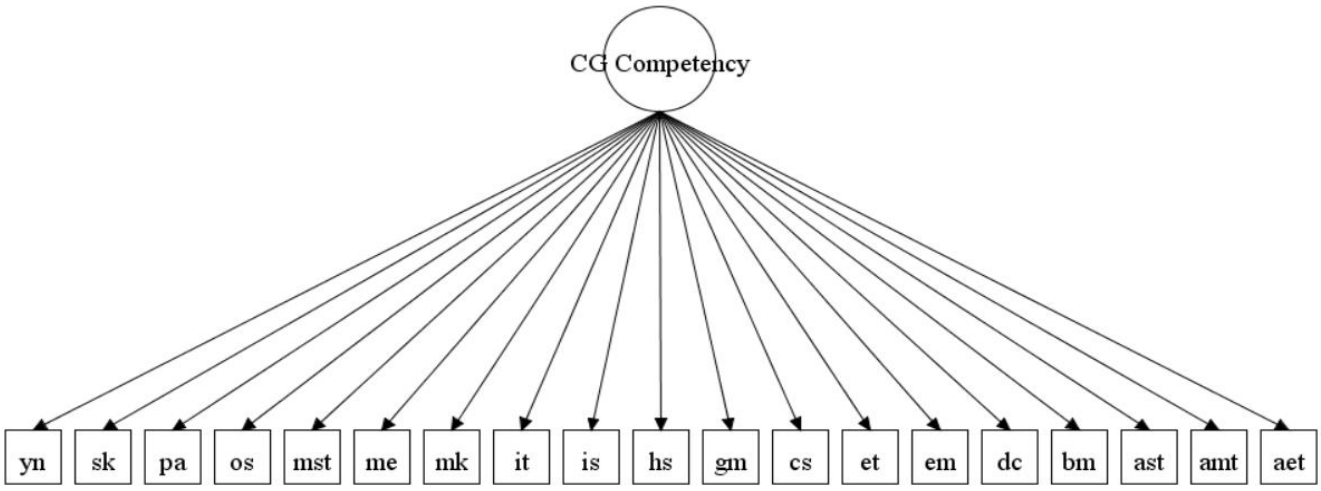
Maintenance Technician [AMT] is 20 weeks), a recruit is considered to be recycled when they spent more time in the training school than originally intended. To calculate this dichotomous outcome (1 = "recycled"; 0 = "not recycled"), we first calculate the time that each USCG recruit spent in school, that is the interval between course start and course end dates, and we then compared it to the standard training length of the school they attended. If days spent in school exceeds the standard training length, the recruit is considered to be recycled. To verify that this definition is not detecting trivial amounts of time beyond the standard training length (e.g., due to a federal holiday during a weekday that extended the training by one day), 93% of recruits we classified as "recycled" attended their training for 8 or more days at their training school.

### *Coast Guard (CG) Competency Score*

We also created dichotomous qualification indicators for each of 19 Coast Guard 'A' schools that have entrance standards, which indicates whether a person was qualified for a specific school. The qualification indicators (1 = "qualified"; 0 = "not qualified") were determined by comparing the actual ASVAB composite and sub-scores of each recruit to the minimum score requirements for each school. For example, the Operational Specialist (OS) school requires VE+AR ≥ 105. If a recruit's score meets this requirement, they will be marked as "qualified" for the OS school.

With the 19 dichotomous qualification indicators, we further calculated a latent USCG competency score for each person via a unidimensional Rasch measurement model. The IV (Investigator) school was not included in the model because this school has no entrance standard and the corresponding indicator has no variance (i.e., all recruits were qualified). The Rasch model is particularly useful here because it can generate an optimally weighted competency score, while simultaneously modeling the difficulty of admission to each training school. Figure 5 below displays the configuration of the model. The USCG competency score was a composite outcome variable used for later validity analyses, and we have found this score was highly reliable (Cronbach's alpha= .95).

Figure 5.
*Measurement Model of USCG Competency Score*



In general, a higher USCG competency score indicates a person is more qualified for different USCG schools and often has more choices to select the school they will attend. With this Rasch model, we also estimated a difficulty parameter for each school. We saved the person-specific competency scores in the file of "*CG Competency.csv*", and the school-specific difficulty estimates ordered from low to high is shown in Table 11. Both person competency and school difficulty scores are along a logit (log odds unit) scale. As can be seen, schools have various difficulty levels to enter. For example, the Marine Science Tech (*diff* = 4.02) is the most difficult school to enter, whereas Boatswain's Mate (*diff* = -4.73) is the easiest (besides the Investigator school of course, for which all recruits qualify).

Table 11.
*Rasch Difficulty Estimate for Each Training School*

| **School Names** | **IRT Difficulty Estimate** |
|---|---|
| Investigator (IV) | N/A |
| Boatswain's Mate (BM) | -4.73 |
| Maritime Enforcement (ME) | -4.73 |
| Machinery Technician (MK) | -3.2 |
| Health Services (HS) | -1.71 |
| Gunner's Mate (GM) | -1.57 |
| Culinary Specialist (CS) | -1.29 |
| Operations Specialist (OS) | -1.29 |
| Yeoman (YN) | -1.29 |
| Storekeeper (SK) | -0.47 |
| Damage Controlman (DC) | -0.42 |
| Electrician's Mate (EM) | -0.39 |
| Intelligence Specialist (IS) | 0.74 |
| Public Affairs (PA) | 1.79 |
| Aviation Survival Tech (AST) | 2.05 |
| Aviation Maintenance Tech (AMT) | 2.39 |
| Avionics Electrical Tech (AET) | 3.36 |
| Electronics Technician (ET) | 3.36 |
| Information Systems Tech (IT) | 3.36 |
| Marine Science Tech (MST) | 4.02 |

**Descriptive Patterns of ASVAB**

The ASVAB test was originally designed to be taken only once by each recruit, but about 8% of the USCG recruits have chosen to take the test repeatedly. Specifically, 6.8% of the recruits took the test twice; 0.7% took it three times; 0.1% took the test four times, and fewer than 0.1% of participants took the ASVAB more than four times. Table 12 shows the percentages of recruits with each number of ASVAB attempts by recruits' gender, race, and Hispanic identities. A higher percentage of female recruits (9.2%) took the ASVAB repeatedly than male recruits (7.4%). Comparing to the White group (7%), all non-White racial groups had higher percentages

of recruits with multiple ASVAB attempts. The Hispanic group (10%) also has a higher proportion of recruits with multiple ASVAB attempts than the non-Hispanic group (7.2%). These descriptive findings indicate that the minority groups, who may be historically under-represented the US Coast Guard, usually choose to take the ASVAB more times in order to enter training schools.

Table 12.
*Number of Attempts for the ASVAB by Self-Reported Gender, Race and Hispanic Identity*

| **Variables** | **Groups** | **N** | **Number of Attempts** | | | |
|---|---|---|---|---|---|---|
| | | | **1 (%)** | **2 (%)** | **3 (%)** | **4 (%)** |
| *Gender* | | | | | | |
| | Male | 15,262 | 92.6 | 6.8 | 0.6 | 0.1 |
| | Female | 2,948 | 90.8 | 7.8 | 1.2 | 0.2 |
| *Race* | | | | | | |
| | White | 13,462 | 93.0 | 6.4 | 0.5 | 0.1 |
| | Black or African American | 1,094 | 85.6 | 11.8 | 2.3 | 0.4 |
| | Asian | 398 | 90.7 | 8.3 | 1.0 | 0.0 |
| | American Indian/Alaska Native | 353 | 89.5 | 9.6 | 0.6 | 0.3 |
| | Native Hawaiian/Pacific Islander | 267 | 89.1 | 10.1 | 0.7 | 0.0 |
| | Multiple Races | 526 | 93.9 | 5.7 | 0.4 | 0.0 |
| | No records (missing data) | 2,110 | 91.9 | 6.9 | 1.0 | 0.2 |
| *Hispanic or Latinx* | | | | | | |
| | No | 12,607 | 92.8 | 6.6 | 0.5 | 0.1 |
| | Yes | 3,191 | 90.0 | 8.8 | 1.1 | 0.1 |
| | No records (missing data) | 2,412 | 92.9 | 6.1 | 1.0 | 0.1 |

As an initial descriptive step to understand the way recruits improved when they took the ASVAB multiple times, we also calculated and compared the magnitude of improvement in ASVAB scores when recruits took the test repeatedly across demographic groups (see Table 13). For the sub-sample ($n$ = 1,403) who had multiple ASVAB attempts, we calculated the magnitude of improvement by subtracting their 1st ASVAB score from the final score. The mean improvement of all these cases was 5.49 ($SD$ = 9.19), and the magnitude varied across demographic groups. Specifically, the female group ($M$ = 6.57, $SD$ = 9.14) had a larger score increase than the male group ($M$ = 5.21, $SD$ = 9.17); the Hispanic group ($M$ = 6.58, $SD$ = 10.44) showed larger improvement than the non-Hispanic group ($M$ = 4.89, $SD$ = 8.74); and all racial minority groups improved more than the White group ($M$ = 5.04, $SD$ = 9.15) on average.

Table 13.
*ASVAB Improvement by Self-Reported Gender, Race and Hispanic Identity*

| Variables | Groups | N | Mean | SD |
|---|---|---|---|---|
| *Gender* | | | | |
| | Male | 1,131 | 5.21 | 9.17 |
| | Female | 270 | 6.57 | 9.14 |
| *Race* | | | | |
| | White | 939 | 5.04 | 9.15 |
| | Black or African American | 157 | 5.36 | 9.20 |
| | Asian | 37 | 7.27 | 8.18 |
| | American Indian/Alaska Native | 37 | 5.22 | 8.07 |
| | Native Hawaiian/Pacific Islander | 29 | 8.48 | 9.44 |
| | Multi Race | 32 | 7.69 | 9.32 |
| *Hispanic or Latinx* | | | | |
| | No | 912 | 4.89 | 8.74 |
| | Yes | 318 | 6.58 | 10.44 |

This descriptive analysis showed that, in general, participants with minority identities had a larger score increase and may benefit more by taking the ASVAB multiple times. Therefore, allowing recruits to take the ASVAB repeatedly could potentially increase the diversity in USCG recruitment. The observed growth also indicates that a single-timepoint ASVAB score might not validly represent recruits' learning potential, and people might underestimate their future performance by only referencing the single ASVAB score. In the sections to come, we apply a statistically complex dynamic measurement model (DMM) in order to more fully investigate this possibility.

To test this more rigorously, we fit regression models to compare differences in the overall change among different demographic groups. To keep sample sizes high and given the USCG's interest in diversity in general rather than a particular demographic group, we collapsed the racial identities into White and non-white. Because their were several outliers in the data cause by some recruits would make large gains in their score (e.g., one recruit increase from a score of 16 to a score of 79 across ASVAB attempts), we used robust regression with M-estimation (Huber, 1973) and a bisquare weighting (Hampel et al., 1986) in SAS PROC ROBUSTREG to test for differences in growth rates. The difference in growth rates for non-White ($\chi^2(1) = 9.37, p = .002$), Hispanic ($\chi^2(1) = 11.23, p < .001$), and female ($\chi^2(1) = 6.82, p = .009$) recruits were all positive and statistically significant, indicating that recruits from underrepresented group improve more that historical majority groups when they take the ASVAB multiple times.

## Section B.2: Building A Dynamic Measurement Model

A straightforward strategy to handle some recruits taking the ASVAB multiple times is to use the highest ASVAB score from each recruit. This may be sufficient for the training school assignment function of the ASVAB, but it discards some potentially useful information about *who* is self-selecting into taking the ASVAB multiple times, which may provide insight for improving the consequential validity of the ASVAB for assigning recruits to training schools. This is relevant to the recent USCG initiatives, which have called for maintaining and increasing a diverse and inclusive workforce (e.g., https://www.dcms.uscg.mil/Portals/10/CG-1/diversity/DIAP/Diversity-and-Inclusion-Action-Plan.pdf?ver=2020-06-25-153724-670). Because training assignments occur early in recruits' careers, this process has the potential to have long-lasting implications on careers for coastguardsmen and the eventual demographic diversity of different roles and ranks.

A major goal of this project was therefore to develop a suitable statistical model for self-selecting into multiple ASVAB attempts. Part of the complexity of these data is that – unlike common personnel selection or licensure exam settings – there is no single score for which an arbitrary recruit is striving. That is, there are many different training schools, each with unique entrance standards, and there also may be idiosyncratic differences for why someone may want to take the ASVAB multiple times. This makes the recent statistical literature on modeling development focused on a single criterion score ineffective for this specific problem (e.g., Johnson & Hancock, 2019; Preacher & Hancock, 2015). Additionally, the data structure is not consistent across all people. A majority of people have only one ASVAB score whereas a small – but nontrivial – number of people have repeated measures for the ASVAB. The self-selection in taking the ASVAB multiple times violates assumptions of most "off-the-shelf" statistical models and required creation of a bespoke statistical model tailored specifically to the nuanced characteristics of these data.

### Building a Model for Self-Selecting into Multiple ASVAB Administrations

To model the unique aspects of these data, we created a statistical model that contains three submodels. Submodel 1 is a growth model for ASVAB scores (i.e., how do ASVAB scores change over time for recruits with multiple scores?). Submodel 2 is a latent class model for self-selection into taking the ASVAB multiple times (i.e., are there variables that can predict who will select to take the ASVAB multiple times?). Submodel 3 is a discrete-time survival model for how many times recruits take the ASVAB (i.e., of the recruits who take the ASVAB multiple times, are there variables that explain how many times a recruit will continue to take the ASVAB?).

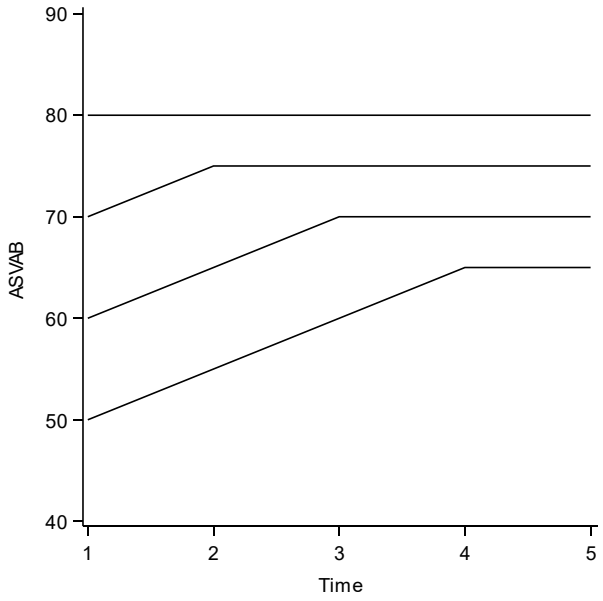
The ASVAB growth submodel consists of a bilinear spline using definition variables for recruit-specific measurement occasions. The submodel for selection into multiple ASVAB administrations is a known-class mixture model for partial clustering to differentiate the model for recruits with and without multiple ASVAB attempts. The submodel for how many times the ASVAB is taken will be a Wu-Carroll shared parameter selection model for informative dropout (Wu & Carroll, 1988). We overview each one of these aspects individually prior to presenting the details of the complete model.

**ASVAB Growth Submodel**

Because the highest ASVAB score is eventually used for placing USCG recruits, we use a model that is parameterized in terms of growth rate and ASVAB Capacity rather than the more traditional intercept and slope parameterization. This is also consistent with the dynamic measurement paradigm discussed earlier in this report. Specifically, we use a bilinear spline that has linear growth prior to the knot point, but growth after the knot point has a slope of zero (i.e., the trajectory after the knot point is a horizontal line).

The idea behind this bilinear spline is that the knot point represents the point where the maximum score has been achieved, so the value of the outcome at the knot point carries forward for the remainder of the observation window. Figure 6 shows a conceptual plot of the growth trajectory for hypothetical people with different numbers of ASVAB scores. The growth rate and the capacity are recruit-specific and allowed to have individual attempts. Because these data have a relatively small number of repeated measures, the transition from the growth phase to the capacity phase is more abrupt than the conceptual figure presented in Figure 1 because fewer ASVAB scores do not provide enough granularity to fit a smooth curve with precision.

Figure 6.
*Conceptual Plot of Growth Model for ASVAB Scores for Recruits With a Different Number of Hypothetical Administrations of the ASVAB*




*Note.* The knot point is the transition from the growth phase to the recruit reaching their capacity. Once capacity is reached, there is no longer an increase in scores and the trajectory transitions to a horizontal line.

This idea is related to floor-ceiling splines for minimum and maximum values (Feng et al., 2019) and to growth offset models for developmental processes with maximum values (McNeish et al., 2021). Because we have few measurement occasions on most recruits, our approach deviates from existing models slightly. First, because many of the recruits have few data points, the

options for modeling the growth trajectory are limited. Despite methodological work highlighting intriguing possibilities for splines (e.g., Harring et al., 2021), our data structure essentially is limited to linear growth only given that so few recruits have four or more timepoints necessary to consider growth trajectories with more nuance. This designates our trajectory as a "broken stick" spline (e.g., Long & Ryoo, 2010) because the trajectories connect before and after the knot point, but the first derivatives are not guaranteed to be equal on either side of the knot point (Cudeck & Klebe, 2002). Our model differs from the classic broken stick spline in that we assume no growth (i.e., slope of zero) after the knot point to map onto the idea of capacity as defined by dynamic measurement models.

Second, existing models typically estimate the knot point as a model parameter. Given our data structure and the fact that many recruits only have one point, we instead code *Time* specific to each recruit such that *Time* = 0 corresponds to the last measurement. If there are multiple ASVAB attempts, *Time* is coded negatively such that the repeated measure before dropout is -1, two repeated measures before dropout is -2, and three repeated measures before dropout is -3. In this way, the intercept corresponds to the recruit-specific knot point such that the model directly estimates the capacity (on the scale of the ASVAB scores) rather than the predicted value as baseline (e.g., Biesanz et al., 2004), although this distinction is irrelevant for people with only one ASVAB score. Similarly, the intercept variance captures between-recruit variance in the capacity rather than the between-recruit variance at baseline. This also allows the model to adapt to recruits with only a single ASVAB score because the knot point is just set to the first test administration such that the model reduces to a random intercepts-only model and the trajectory is a horizontal line (i.e., without additional information, the capacity is assumed to be equal to the only observed datapoint).

Given the multivariate nature of our model, we operate in the latent growth framework rather than the multilevel framework (MacCallum et al., 1997). Although latent growth models are desirable for their ability to embed within larger multivariate models, recruit-specific measurement occasions are more challenging to accommodate in this framework (e.g., McNeish & Matta, 2018). This difference arises because *Time* is not an explicit predictor in a latent growth model; instead, *Time* is implied via constraints on the basis coefficients from the slope growth factor to the repeated measures. Coding *Time* such that 0 corresponds to the last ASVAB administration to make the intercept interpretable as the capacity results in recruit-specific values of *Time*. For instance, *Time* for recruits with only one ASVAB attempt would be coded as $\begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix}$ whereas a recruit with four ASVAB attempts would have *Time* coded as $\begin{bmatrix}-3 & -2 & -1 & 0\end{bmatrix}$.

We address this issue using *definition variables* (Mehta & West, 2000; Mehta & Neale, 2005). Definition variables constrain basis coefficients to a recruit-specific variable in the dataset, rather than to a constant value (Grimm & Ram, 2009), which more closely aligns with how a design matrix is built in a multilevel model where *Time* is an explicit predictor and recruit-specific values of *Time* are not problematic (Blozis & Cho, 2008). This approach has also been discussed in the context of spline models specifically when 0 corresponds to a knot point (Sterba, 2014).

**Known-Class Mixture Models and Partial Clustering**

The structure of the data where some recruits have only one ASVAB attempt while others have multiple ASVAB attempts has been referred to as *partial clustering* (Sterba et al., 2014). More commonly, this type of data structure is seen in clinical trials (Sterba, 2017) such that data from the control arm is independent but data from the experimental arm is clustered (Bauer et al., 2008). For instance, an intervention might aim to compare the effectiveness of group therapy; people who are assigned to the experimental arm would be assigned to therapy group with other experimental arm participants (and people are therefore clustered with therapy groups), but control arm participants would be independent (Baldwin et al., 2011). Though less common for repeated measures, the same principle applies if some people only have a single datapoint, but others have repeated measures. Data from recruits with a single datapoint are independent but data from recruits with repeated measures are clustered, making the overall data structure partially clustered. In the USCG data, rather than being randomly assigned to these conditions as in the typical partially clustered design, recruits have the ability to self-select into having repeated measures.

If ignoring the partially clustered design and fitting a typical growth model to all recruits from a partially clustered data structure, recruits without repeated measures should have no possibility of growth and the random slope distribution would be a point mass at 0. This would have ramifications for the slope variance and would severely underestimate the between-recruit variability in growth for recruits who did have repeated measures. Alternatively, it would be possible to assume a recruit-specific slope for recruits with only one repeated measure, but the empirical Bayes prediction for the slope would be heavily shrunken to the point of essentially being equal to the fixed effect. Furthermore, the random slope distribution would be unlikely to be normal, especially considering that a sizeable number of recruits only have one ASVAB attempt. This is problematic for the dropout selection model discussed in the next subsection because the model is sensitive to distributional assumptions, so such a clear violation would be problematic.

In these situations, Kim et al. (2014) suggest known class mixture models as a way to model covariance structures that vary for different types of recruits in the data. This allows recruits with multiple ASVAB attempts to be in a separate class from recruits with a single ASVAB attempt. It also allows the growth model to be different in each class to reflect that one of the classes has independent data whereas the other class has repeated measures. This type of model is related to zero-inflated and one-inflated models used in criminology, substance use, and demography whereby there is a large spike in data at extreme low values of a variable (Ospina & Ferrari, 2012; Liu & Eugenio, 2018). The difference in this current project, however, is that the one-inflatedness of the data concerns the number of repeated measures rather than the one-inflatedness of the outcome variable itself.

Known class mixture models are conceptually related to multiple group models but are more flexible with respect to estimation. Specifically, in the M*plus* software, models with categorical outcomes (which will be presented in the dropout selection model subsection below) and continuous latent variables (from the growth model) result in a likelihood that does not have a closed form and the integral in the likelihood must be solved numerically (Muthén & Muthén, 1998-2022). This cannot be accomplished in a multiple group model but can be completed with a

mixture model. Therefore, the known class mixture model specifies the model with latent classes, but class assignment is done deterministically as a function of an observed variable rather than probabilistically as in a standard mixture model.

The known class mixture model has the advantage that it creates a discrete latent variable corresponding to whether recruits had only one ASVAB attempt or whether they had multiple ASVAB attempts. A logistic regression submodel can also be included that uses the discrete latent class as the outcome such that covariates can be included to model selection into ASVAB attempts. This can help address one of the main research questions regarding characteristics of recruits that are self-selected into multiple ASVAB attempts.

**Selection Models for Informative Dropout**

Within the latent class corresponding to recruits who self-selected into multiple ASVAB attempts, there is also a self-selection process into how many attempts recruits decide to take. This type of process creates data that are missing not at random (MNAR; Rubin, 1976) such that the reason for the value being missing is related to what the data value would have been (Enders, 2011; Hedeker & Gibbons, 2006). That is, recruits with higher scores or larger growth between scores are more likely to have fewer ASVAB attempts because they are more likely to decide not to continue to take the ASVAB because their additional ASVAB attempts would have been high had they been collected (i.e., recruits with high ASVAB scores have less incentive to select into additional ASVAB attempts).

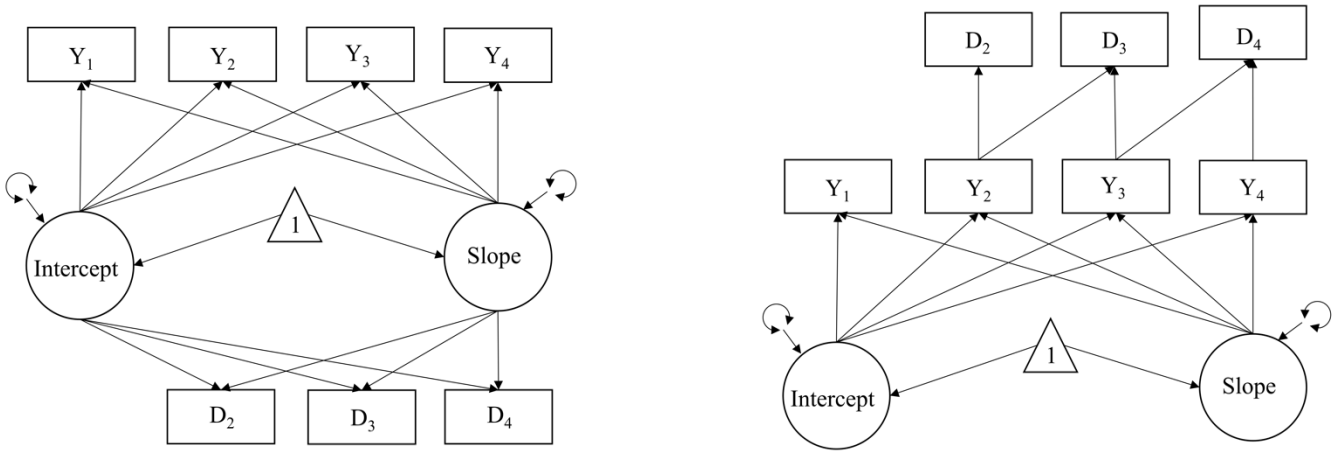
This type of missing data mechanism is relatively uncommon in behavioral research (c.f. Matta & Soland, 2019; Muthén & Masyn, 2005) but occurs quite frequently in medical research. In medical research, patients often dropout of studies and have MNAR data because they are too sick to continue or because they recover from their illness and are too healthy to remain in the study (e.g., Thiébaut et al., 2005). Although not necessarily the same mechanism as the self-selection process in which we are interested, the underlying statistical principle is the same. That is, patients have missing values precisely because of the values that they would have provided (i.e., dropout is informative). For instance, in a blood cancer trial, a patient with an extremely low red blood cell count may drop out due to severe illness or death. The missing values in this patient's data vector are MNAR because they are directly related to what the values would have been had the data been collected.

A common class of methods from the biostatistics literature by which to model MNAR data in longitudinal studies is joint modeling and selection models (Tsiatis & Davidian, 2004). The idea of these models is to jointly model the change in the outcome variable with a growth model and the dropout mechanism with a survival model. That is, in addition to the primary outcome of the study (ASVAB score), a binary survival indicator is created such that "0" means that the person remains in the study and "1" means that the person has dropped out. As in traditional survival analyses, researchers can predict dropout as a function of covariates that are hypothesized to be related to why people would discontinue participating in the study (typically with a logistic regression). These covariates can be exogenous covariates like demographic characteristics or can be variables from the growth process such as observed repeated measures or random intercepts and slopes. The idea is that if researchers can model the selection factors that are

causing dropout, then the missingness is no longer MNAR and the model's inferences and conclusions can be trusted.

The two major types of selection models are differentiated by what predicts the binary survival indicators. The Wu-Carroll shared parameter model (Wu & Carroll, 1988) uses recruit-specific intercepts and slopes to predict missing values whereas the Diggle-Kenward selection model (Diggle & Kenward, 1994) uses the observed repeated measures from the current and previous time point. Figure 7 shows two hypothetical latent growth model path diagrams to differentiate each model.

Figure 7.
*Comparison of Wu-Carroll Selection Model (Left Panel) and Diggle-Kenward Selection Model (Right Panel)*




*Note*. "Y" variables are repeated measures, "D" variables are binary survival indicators of dropout. The Wu-Carroll model predicts the survival indicators based on the growth factors (plus additional covariates, if desired) whereas the Diggle-Kenward model predictors the survival indicators based on the values of the observed repeated measures (plus additional covariates, if desired)

The Wu-Carroll model is useful for modeling missingness as a function of an entire growth trajectory, which can have advantages over using individual (and possibly error-prone) observed scores (Albert & Follmann, 2000; Little, 1995). The Kenward-Diggle model is useful when missingness is hypothesized to be triggered by a particular threshold of the outcome rather than the general trajectory, as might be the case if there were a entrance standard score for which all recruits were shooting. Either model is heavily influenced by distributional assumptions, which are necessary to permit estimation of the models. In the Wu-Carroll model, the distributional assumptions of the random intercepts and slopes are most relevant; in the Kenward-Diggle model, the distributional assumptions of the outcome themselves (including missing values) are most relevant. Normality is a typical assumption, but it is untestable given that the some of the values are unknown (Kenward, 1998).

In our model, we include a Wu-Carroll selection model to model dropout in the latent class composed of recruits with multiple ASVAB administrations. We elected to use the Wu-Carroll model rather than the Kenward-Diggle model because we did not anticipate that dropout would be triggered by particular ASVAB values given that there are 20 different training schools that each have different entrance standards. We also included demographic characteristics as covariates in the selection model given the CG's dedication to maintaining and expanding an inclusive workforce. The next section provides complete detail about combining all three submodels into one full statistical model.

## The Full Model

The full model is shown in Equation 1.

$$\begin{aligned} \mathbf{y}_{ik} &= \mathbf{\Lambda}_{ik}\boldsymbol{\eta}_{ik} + \boldsymbol{\varepsilon}_{ik} \\ \boldsymbol{\eta}_{ik} &= \boldsymbol{\alpha}_{k} + \boldsymbol{\zeta}_{ik} \\ \ln\left(\frac{\Pr(C_i = 1)}{1 - \Pr(C_i = 1)}\right) &= \alpha_C + \mathbf{X}_i\boldsymbol{\omega} \end{aligned} \tag{1}$$

The first expression shows that the vector of (potentially repeated) ASVAB scores $\mathbf{y}_{ik}$ for recruit *i* in class *k* is modeled as a matrix of basis coefficients $\mathbf{\Lambda}_{ik}$ that are both recruit- and class-specific multiplied by recruit- and class-specific growth factors $\boldsymbol{\eta}_{ik}$ plus a vector of recruit- and class-specific within-recruit residuals $\boldsymbol{\varepsilon}$. The basis coefficient matrix $\mathbf{\Lambda}_{ik}$ in latent growth models does not typically have an *i* subscript, but it is included here because we are incorporating definition variables to permit each recruit's basis coefficients to be potentially different. The second expression then shows that the vector growth factors $\boldsymbol{\eta}_{ik}$ is modeled as a vector of class-specific growth factor means $\boldsymbol{\alpha}_{k}$ plus a vector of recruit- and class-specific disturbances $\boldsymbol{\zeta}_{ik}$.

The third expression in Equation 1 is a logistic regression to model self-selection into latent classes ($C_i$) where $C_i = 0$ corresponds to recruits with only one ASVAB score and $C_i = 1$ corresponds to recruits with multiple ASVAB scores. $\alpha_C$ corresponds to the log-odds of selecting into Class 1 when all covariates are 0, $\mathbf{X}_i$ is a row vector of recruit *i*'s covariates that predict class membership, and $\boldsymbol{\omega}$ is a column vector of covariate effects capturing the change in log-odds of being in Class 1. In our model, $\mathbf{X}_i$ contains three demographic characteristics: self-reported ethnicity (Hispanic =1, Non-Hispanic=0), self-reported sex (Female =1, Male =0), self-reported race (Non-White =1, White = 0), and all two- and three-way interaction between them. Race was collapsed into two categories because the frequencies of specific categories were not high enough to model specific effects of each category individually.

The structure of the vectors and matrices with a *k* subscript vary depending on whether the recruit is in Class 0 or Class 1. If the recruit has one ASVAB attempt,

$$\text{if } C_i = 0 \begin{cases} \boldsymbol{\Lambda}_k = [1] \\ \boldsymbol{\eta}_{ik} = [\eta_{0i}] \\ \boldsymbol{\alpha}_k = [\alpha_0] \\ \zeta_{0ik} \sim N(0, \psi_{00}) \\ \varepsilon_{ik} \sim N(0, \theta) \end{cases} \tag{2}$$

When $C_i = 0$, data are independent so there is no information related to change over time. Each recruit's ASVAB score is modeled as the average ASVAB for Class 0 ($\alpha_0$) plus a recruit-specific disturbance ($\zeta_{0ik}$) that captures the deviation of recruit *i*'s ASVAB from the Class 0 average. This disturbance is assumed to be normally distributed, as is the within-recruit residual variance. This model is a random intercept-only model (e.g., Grimm et al., 2016) such that everyone's trajectory is a horizontal line given that everyone in Class 0 selected to stop taking the ASVAB after one administration.

If a recruit has more than one ASVAB attempts,

$$\text{if } C_i = 1 \begin{cases} \boldsymbol{\Lambda}_{ik} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ a_{1i} & a_{2i} & a_{3i} & a_{4i} \end{bmatrix}' \\ \boldsymbol{\eta}_{ik} = \begin{bmatrix} \eta_{0i} & \eta_{1i} \end{bmatrix}' \\ \boldsymbol{\alpha}_k = \begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix}' \\ \boldsymbol{\varepsilon}_{ik} \sim MVN\left( \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \theta & & & \\ & \theta & & \\ & & \theta & \\ & & & \theta \end{bmatrix} \right) \\ \boldsymbol{\zeta}_{ik} \sim MVN\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \psi_{11} & \\ \psi_{21} & \psi_{22} \end{bmatrix} \right) \end{cases} \tag{3}$$

In Equation 3, notice that the basis coefficients associated with the slope in $\boldsymbol{\Lambda}$ are recruit-specific definition variables, not constants. The values of *a* depend on the number of repeated measures collected for recruit *i*. For two repeated measures, $\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 \end{bmatrix}'$. The basis coefficient matrix indicates that there was growth between Time 1 and Time 2 and the recruit reached their capacity at Time 2. That capacity is then carried forward, without growth. Similarly, with three

repeated measures $\mathbf{\Lambda} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 0 \end{bmatrix}'$ to indicate that there was growth from Time 1 and Time 2 and from Time 2 to Time 3 but that the capacity was reached at Time 3 and this value is carried forward. Lastly, with four repeated measures $\mathbf{\Lambda} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -3 & -2 & -1 & 0 \end{bmatrix}'$ and the change in ASVAB scores is modeled linearly from Time 1 to Time 4. The basis coefficients are modeled this way to parameterize $\eta_{0i}$ as the capacity rather than the ASVAB score at the first attempt if *Time* where implied in the traditional manner of coding the first observation as 0 and counting up.

Equation 3 also has two growth factor means corresponding to the average capacity in Class 1 ( $\alpha_1$ ) and linear growth in ASVAB scores for each additional test attempt ( $\alpha_2$ ). The recruit-specific disturbances capturing each recruit's deviation from the Class 1 average follow a multivariate normal distribution that allow the disturbances to covary. Note that the subscripts in Equation 3 are distinct from the subscripts in Equation 2 because the growth factor means and disturbance covariance parameters are uniquely estimated in each class. The within-recruit residuals in Class 1 are assumed to follow a multivariate normal distribution whose variances are constrained across time and across classes to facilitate estimation given the varying number of attempts per recruit.

The Class 1 model also features a Wu-Carroll shared parameter selection model for when recruits select to stop taking the ASVAB.

$$\begin{aligned} \ln\left(\frac{\Pr\left(D_{3i}=1\right)}{1-\Pr\left(D_{3i}=1\right)}\right) &= \tau_{D3} + \mathbf{X}_i\boldsymbol{\gamma} + \boldsymbol{\beta}\boldsymbol{\eta}_{ik} \\ \ln\left(\frac{\Pr\left(D_{4i}=1\right)}{1-\Pr\left(D_{4i}=1\right)}\right) &= \tau_{D4} + \mathbf{X}_i\boldsymbol{\gamma} + \boldsymbol{\beta}\boldsymbol{\eta}_{ik} \end{aligned} \tag{4}$$

Survival indicators $D_{3i}$ and $D_{4i}$ are created *t*o capture whether recruit *i* had selected to stop taking the ASVAB at Time 3 and Time 4, respectively. Following suggestions in Enders (2005), if recruit *i* was still taking the ASVAB at Time 3, then $D_{3i} = 0$ ,if recruit *i* took the ASVAB twice and selected to dropout at Time 3, then $D_{3i} = 1$. The same logic is applicable to $D_{4i}$ with the additional stipulation that recruits who dropped out after two ASVAB attempts have missing values for $D_{4i}$ (i.e., a "1" value indicates dropout specifically at the specified time-point rather than at or before the specified time-point).

The survival indicators are then modeled with a logistic regression. $\tau_{D3}$ and $\tau_{D4}$ are thresholds that capture the opposite of the log-odds of dropping out when all predictors equal zero at Time 3 or Time 4, respectively (i.e., the threshold is negative one multiplied by the intercept in a traditional logistic regression). The log-odds of dropout are further modeled as a function of four

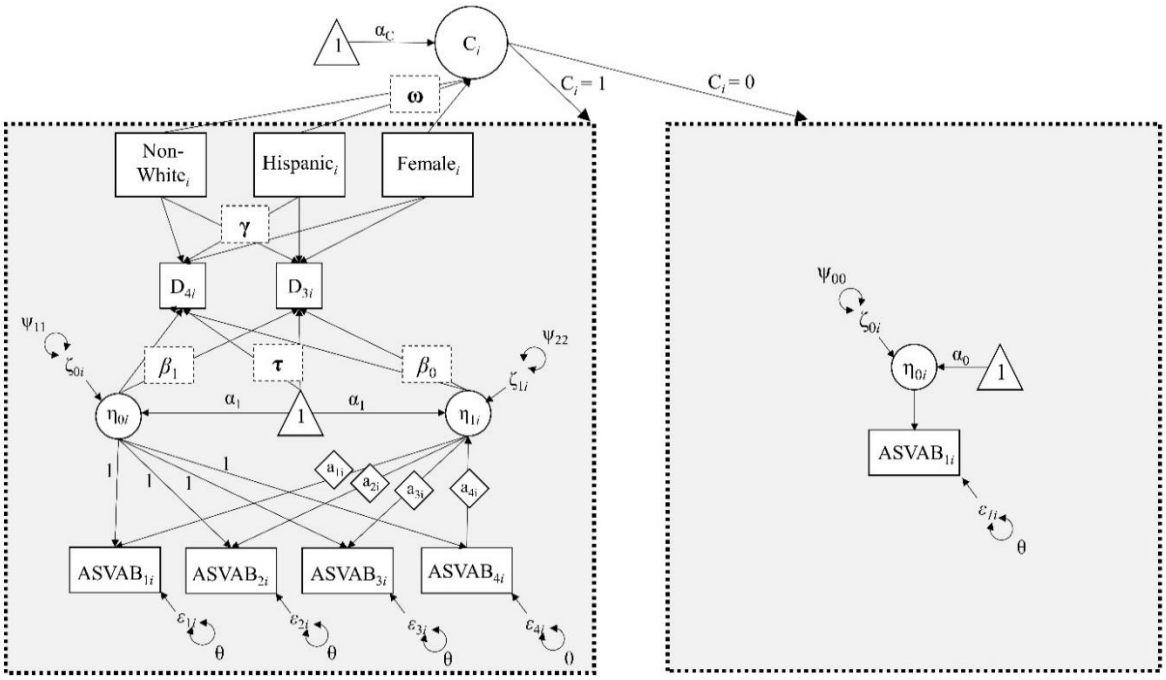
covariates in $\mathbf{X}_i$ whose effects are contained in $\boldsymbol{\gamma}$: self-reported ethnicity (Hispanic =1, Non-Hispanic=0), self-reported sex (Female =1, Male =0), self-reported race (Non-White =1, White = 0), and all two- and three-way interaction between them. These are the same covariates used to predict latent class membership in Equation 1.

These covariate effects in $\boldsymbol{\gamma}$ can be modeled to be unique across time; however, we constrain them to be time-invariant to make estimation more stable, especially because the sample size is small for recruits with four repeated measures. The dropout model also includes the growth factors $\boldsymbol{\eta}_{ik}$ from the ASVAB growth model such that the log-odds of dropout can change depending on a recruit's ASVAB capacity or their growth in ASVAB scores. The effects of the growth factors on dropout are contained in the $\boldsymbol{\beta}$ vector. Similar to effects in $\boldsymbol{\gamma}$, the effect of growth factors on dropout can be modeled as unique at each repeated measure, but we model them as time-invariant to improve numerical stability.

**Conceptual Path Diagram**

A conceptual path diagram of the full model is shown in Figure 8. At the top of the path diagram is the latent class variable *C*, which manifests two different models depending on whether recruit *i* has one ASVAB attempt ($C_i = 0$) or multiple ASVAB attempts ($C_i = 1$). The latent class variable is predicted by four demographic variables to assess whether recruits with different backgrounds are more likely to select into the different classes.

Figure 8.
*Conceptual Path Diagram of Full Dynamic Measurement Model*




*Note*. Circles are latent variables, rectangles are observed variables, triangles are constants, dashed rectangles are parameter vectors, and grey rectangles represent latent classes.

The light grey rectangle on the right shows the $C_i = 0$ model for recruits with one ASVAB attempt. This model is much simpler and is a random intercepts-only model. The intercept variance and intercept growth factor mean both have different subscripts than the corresponding $C_i = 1$ model because these parameters are uniquely estimated in each class.

The light grey rectangle on the left shows the $C_i = 1$ model for recruits with multiple ASVAB attempts. The bottom portion of the diagram shows the broken-stick spline dynamic measurement model for ASVAB capacity. The diamonds placed over the basis coefficients associated with the latent variable for growth ($\eta_{1i}$) indicate that definition variables are used to determine the values of this path uniquely for each recruit. The middle portion of the $C_i = 1$ model shows the selection model for recruits who dropout after two ASVAB attempts (when $D_{3i} = 1$) and who dropout after three ASVAB attempts (when $D_{4i} = 1$). Everyone in the $C_i = 1$ model has at least two ASVAB attempts, so there is no $D_{2i}$ because the variable would have no variance. In the selection part of the model, dropout is predicted by the latent variable representing ASVAB capacity ($\eta_{0i}$), the latent variable for growth ($\eta_{1i}$), and the demographic covariates (which have three main effect, three two-way interactions, and one three-way interaction).

## Section B.3: Dynamic Measurement Modeling Results

### Model Fitting

We fit the model in M*plus* Version 8.7, which supports (a) the known-class model with a `KNOWNCLASS` option in the `VARIABLE` statement, (b) the definition variable approach with the `TSCORES` option in the `VARIABLE` statement, and (c) creating survival indicators using the `TYPE=SDROPOUT` option in the `DATA MISSING` statement. Model parameters were estimated with robust maximum likelihood with adaptive Gaussian quadrature using 15 quadrature of integration for both endogenous latent variables ($\eta_{0i}$ and $\eta_{1i}$), resulting is 225 points of integration per iteration of the optimization algorithm.

### Growth Submodel Results

The parameter estimates, *p*-values[1], and odds ratios (where applicable) are shown in Table 14, which separates the parameters into three sections to delineate the growth model, the class membership model, and the dropout selection submodels. The ASVAB growth model is different in each class, the class membership model is constant across all classes, and the dropout selection model only exists in the Multiple Attempt class. There are no available model global fit criteria for this model because the dimension of the covariance structure is recruit-specific.

In the ASVAB growth model, the results show that a recruit who selects into multiple ASVAB attempts times grows about 5 points, on average, with each additional attempt ($\alpha_2 = 5.04, Z = 21.40, p < .01$). Using a one-sided test because variances cannot be negative (Liu, 1997), there was significant between-recruit variability in the predicted growth ($\psi_{22} = 6.26, Z = 1.89, p_{one-tail} = .03$). Assuming within-class normality in the Multiple Attempt class, 95% of recruit-specific growth rates would therefore fall between 0.14 and 9.94 points per additional ASVAB attempt.

[1] Even though the data contain all ASVAB scores for years May 2013 to June 2021, we still report *p*-values because this is not a historical analysis, and we consider future CG recruits as part of the population of interest. Based on this definition, we do not have the entire population and statistical inference is still necessary.

Table 14.
*Parameter estimates for full model*

| Model | Parameter | $C_i = 1$ Estimate | *p*-value | $C_i = 0$ Estimate | *p*-value |
|---|---|---|---|---|---|
| ASVAB Growth | | | | | |
| | Capacity, Mean | 65.73 | <.01 | 72.05 | <.01 |
| | Growth, Mean | 5.04 | <.01 | --- | --- |
| | Capacity, Variance | 188.65 | <.01 | 200.41 | <.01 |
| | Growth, Variance | 6.26 | .03 | --- | --- |
| | Cap., Growth Correlation | -0.07 | .54 | --- | --- |
| | Residual Variance | 38.49 | <.01 | 38.49 | <.01 |
| Class Membership | | | Estimate | p-value | Odds Ratio |
| | Intercept | | -2.73 | <.01 | |
| | Female | | 0.28 | <.01 | 1.33 |
| | Non-White | | 0.44 | <.01 | 1.55 |
| | Hispanic | | 0.49 | <.01 | 1.64 |
| | Female × Non-White | | -0.02 | .92 | 0.98 |
| | Female × Hispanic | | -0.26 | .22 | 0.77 |
| | Non-White × Hispanic | | -0.37 | .03 | 0.69 |
| | Female × Non-White× Hispanic | | 0.17 | .63 | 1.19 |
| Dropout Selection | | | $C_i = 1$ Estimate | p-value | Odds Ratio |
| | Time 3 Threshold | | 2.67 | .13 | |
| | Time 4 Threshold | | 1.71 | .24 | |
| | ASVAB Capacity | | 0.06 | <.01 | 1.07 |
| | ASVAB Growth | | 0.31 | .17 | 1.40 |
| | Female | | -0.70 | .04 | 0.49 |
| | Non-White | | -0.59 | .03 | 0.56 |
| | Hispanic | | -0.52 | .09 | 0.60 |
| | Female × Non-White | | 0.03 | .95 | 1.03 |
| | Female × Hispanic | | 0.43 | .50 | 1.53 |
| | Non-White × Hispanic | | 0.79 | .22 | 2.21 |
| | Female × Non-White× Hispanic | | -0.83 | .42 | 0.44 |

On average, a recruit with multiple ASVAB attempts had a maximum capacity of $\alpha_1 = 65.73$ compared to the average capacity for recruits with a single ASVAB of $\alpha_0 = 72.05$. The between-recruit variance for the capacities between classes was similar but slightly smaller for the Multiple Attempt class ($\psi_{11} = 188.65$ for the Multiple Attempt class vs. $\psi_{00} = 200.41$ for the Single Attempt class), which produces fairly comparable ranges for recruit-specific maximum ASVAB scores: [38.81, 92.65] for the multiple attempt class versus [44.30, 99.80] for the single attempt class such that there is about 83% overlap in the distributions, as presented in Figure 9. In the Multiple Attempt class, the covariance between capacity and the growth rate was not significant ($r = -.07, Z = -0.65, p = .52$), indicating that recruit's capacity was not systematically related to their rate of improvement.

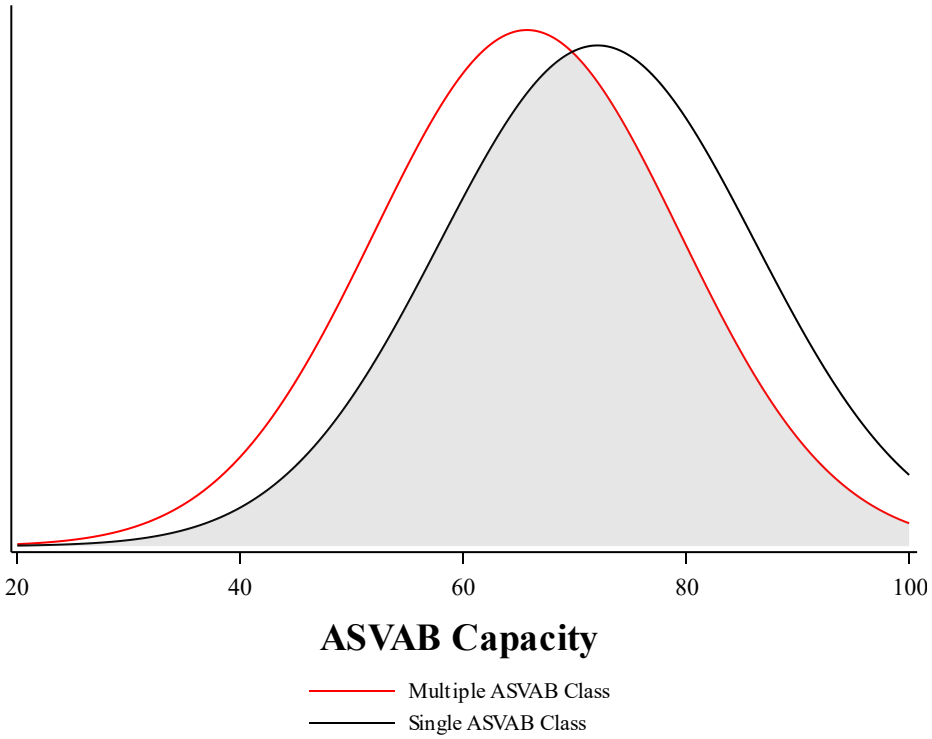


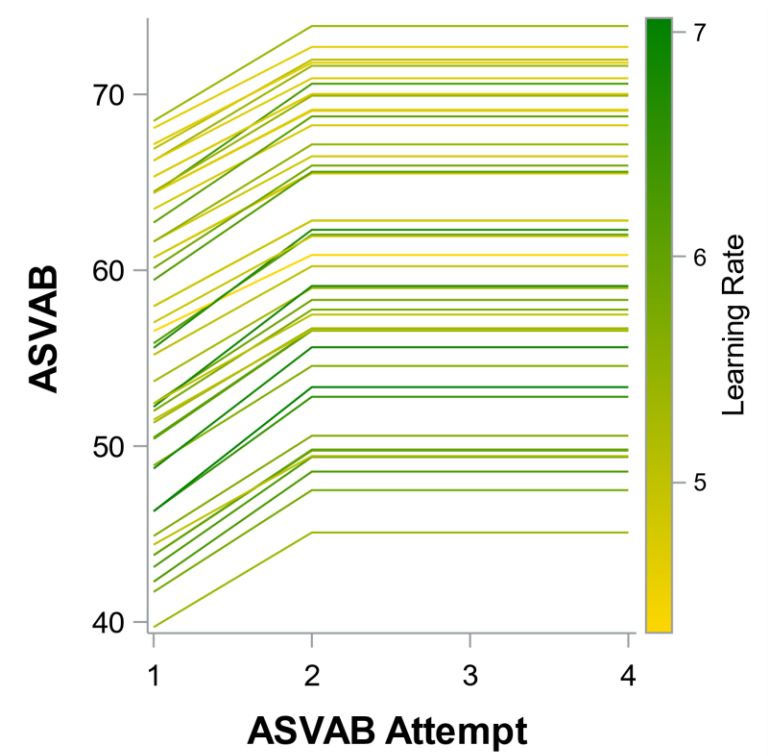

Figure 9.
*Overlap distributions of recruit-specific Maximum ASVAB scores for the class with multiple ASVAB administrations in red (left) and the class with a single ASVAB administration in black (right). The total overlap is 83% and denoted by grey shading.*

Essentially, recruits who elect to attempt the ASVAB multiple times tend to increase their scores. The ASVAB capacity for recruits with multiple attempts is slightly less that the single attempt recruits, but the distributions of ASVAB capacities largely overlap.

To provide a visual representation of the model predicted growth trajectories, Figure 10 shows a gradient plot for a random sample of 50 recruits with 2 ASVAB attempts. In this gradient plot, the color of the line is tied to the growth rate for each recruit. Lines that are dark green represent recruits with more growth between ASVAB attempts and lines that are gold represent recruits

with less growth between ASVAB attempts. After Attempt 2, all the recruits depicted in this plot self-select into no further ASVAB attempts, which is encoded in the model as the recruit having reached their capacity on the ASVAB. Therefore, the growth trajectory transitions to a horizontal line after the second attempt and no longer is permitted to grow.

Figure 10.
*Gradient Plot for Predicted Growth Trajectory for 50 Randomly Sampled Recruits with Two ASVAB Attempts*




*Note*. ASVAB scores can change from Attempt 1 to Attempt 2 and the rate of change is denoted the color of the line (darker green represents more change).

**Class Membership Submodel Results**

With respect to the class membership model for predicting who is in each class, the results show that there are some notable demographic differences in who is self-selecting into multiple ASVAB attempts. First, note that the intercept for membership in the Multiple Attempt class is $\alpha_C = -2.73$ which means that the probability of being in the multiple membership class for a White, Non-Hispanic, Male recruit (i.e., someone who has all 0 values on the covariates) is $\exp(-2.73)/(1+\exp(1-2.68)) = 6.1\%$. This is lower than the 8% of total recruits who have multiple ASVAB scores.

Recruits reporting that they are Female ( $\omega_1 = 0.28, Z = 2.92, p < .01, OR = 1.33$ ), Hispanic ( $\omega_2 = 0.49, Z = 5.60, p < .01, OR = 1.64$ ), or Non-White ( $\omega_3 = 0.44, Z = 5.85, p < .01, OR = 1.55$ ), all had a significantly higher probability of being in the Multiple Attempt class. Among interactions of these characteristics, only the interaction between Hispanic and Non-White was significant and its coefficient was negative, indicating that simultaneously self-reporting Hispanic and Non-White identities lowered the probability of multiple ASVAB attempts relative to the sum of the only self-reporting a Hispanic identify and only self-reporting a Non-White Identity ( $\omega_6 = -0.37, Z = -2.24, p = .03, OR = 0.69$ ).

Table 15 shows the predicted probability of being in the Multiple Attempt class for different demographic combinations. Notably, recruits from groups that are traditionally under-represented in the USCG have a higher probability of taking the ASVAB more often.

Table 15.
*Predicted Probability of Multiple ASVAB Attempts*

| Demographics | Predicted Probability, Multiple Attempts |
|---|---|
| Female, White, & Non-Hispanic | 7.9% |
| Hispanic, White, & Male | 9.6% |
| Non-White, Male, & Non-Hispanic | 9.1% |
| Female, Hispanic, & White | 9.8% |
| Non-White, Hispanic, & Male | 10.2% |
| Female, Non-White, & Non-Hispanic | 11.1% |
| Female, Non-White, & Hispanic | 11.4% |

**Dropout Submodel Results**

The last section of Table 14 above shows the estimates for the selection model for the probability of selecting to stop taking the ASVAB as a function of demographic covariates and latent growth factors. The threshold for Time 3 dropout is $\tau_{3i} = 2.90$ , which means that the predicted probability of a White, Non-Hispanic Male, recruit with an average capacity and an average growth rate ceases to take the ASVAB after two attempts is

$$\begin{aligned} &\exp\left(1+\left(-2.67+65.73\times.064+5.04\times.314\right)\right)\Big/\left(1+\exp\left(1+\left(-2.67+65.73\times.064+5.04\times.314\right)\right)\right) \\ &=\exp\left(3.11\right)\Big/\left(1+\exp\left(3.11\right)\right) \\ &=95.7\% \end{aligned} .$$

Substituting the Time 4 threshold instead yields a probability of 98.3%.

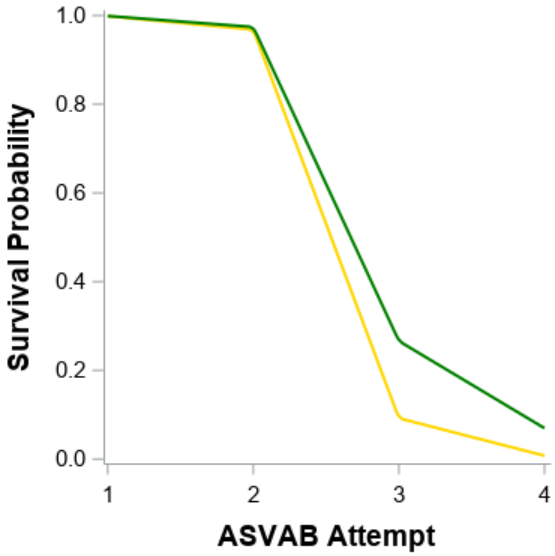
For individuals with larger growth rates between ASVAB attempts, the probability of dropout increases ( $\beta_1 = 0.31, OR = 1.37$ ) although the effect was not significant ( $Z = 1.39, p = .17$ ),

presumably because there was larger uncertainty in growth rate given that only about 10% of the sample were eligible to estimate change over time. Similarly, recruits with higher capacities were significantly more likely to dropout and decide to stop taking the ASVAB ($\beta_0 = .064, OR = 1.07, Z = 4.22, p < .01$).

All main effect coefficients for the demographic indicator variables are negative, indicating that recruits reporting being members of these demographic groups are less likely to dropout, and choose to continue taking the ASVAB. Female and Non- White recruits were particularly less likely to cease taking the ASVAB once they selected into multiple ASVAB attempts. The odds of a female recruit dropping out were about half of a male recruit with an identical capacity and an identical growth rate ($\gamma_1 = -0.70, OR = 0.49, Z = -2.06, p = .04$) and the odds of a Non-White recruit dropping out were about half of a White recruit with identical capacity and growth rate ($\gamma_2 = -0.59, OR = 0.56, Z = -2.25, p = .03$). Other demographic predictors, including all two-way and three-way interactions, were not statistically significant (though keep in mind that the sample sizes for the dropout selection model are much smaller than other portions of the model).

To visualize the difference in the number of ASVAB attempts within the multiple attempt class, Figure 11 shows the survival probabilities for White, Male, Non-Hispanic (in gold) and Non-White, Female, Hispanic recruits (in green). These probabilities are only for recruits in the Multiple Attempt class, so the survival probability at Time 1 and Time 2 is 100% for both groups. At Time 3, only 6% of White, Male, Non-Hispanic recruits continue to take the ASVAB whereas 25% of Non-White, Female, Hispanic recruits continue. Less than half a percent of White, Male, Non-Hispanic recruits attempted the ASVAB a fourth time but over 6% of Non-White, Female, Hispanic recruits attempted the ASVAB a fourth time.

Figure 11.
*Predicted Survival Probability Plot for White, Non-Hispanic, Males (in Gold) and Non-White, Hispanic, Females (in Green).*




*Note.* The predicted probability of three ASVAB attempts is 6% for White, Non-Hispanic, Male recruits represented by the gold line 25% for White, Hispanic, Female recruits represented by the green line.

## Section B.4: Fairness and Validity Analysis of DMM Scores

We performed multiple analyses to examine the fairness and validity of the capacity scores as well as the incremental value of using DMM scores to predict recruits' training outcomes. Validity analyses were all conducted in the Stata software. The codes and output from the analyses are organized in the file named "Validity Analyses Code and Output.pdf".

### What Is the Impact of Demographics on Learning Capacity?

Consequential validity (sometimes called test fairness) appraises the potential and actual consequences of using a test score (Messick, 1989). When test scores are highly associated with the demographics of recruits, negative recruitment consequences (e.g., lack of diversity) may occur. In this section, we first investigated the consequential validity of capacity scores.

We used the demographic variables to predict the saved capacity scores from DMM. Table 16 summarizes the effect sizes of those demographic predictors. It was found that the combination of demographics only explained a small amount of variance ($R^2 = 0.04$) in the capacity scores, and the effect sizes of demographic predictors or their interaction terms were all very small ($\eta^2 < 0.01$). In other words, the capacity scores estimated from DMM were not impacted by recruits' demographics, which provides evidence for good consequential validity of capacity scores.

Table 16.
*Effect Sizes for the Predictive Model of Demographics to Capacity*

| Source | df | $\eta^2$ |
|---|---|---|
| *Model* | 23 | .041 |
| | | |
| Gender | 1 | .000 |
| Hispanic | 1 | .000 |
| Gender × Hispanic | 1 | .000 |
| Non-White | 5 | .003 |
| Gender × Non-White | 5 | .000 |
| Hispanic × Non-White | 5 | .002 |
| Gender × Hispanic × Non-White | 5 | .001 |

*Note*. $\eta^2$ values for individual model terms are partial.

### Can DMM Capacity Scores Support Fair Decisions in Military Recruitment?

To further examine the consequential validity of the capacity scores from DMM, that is whether DMM capacity scores can support fair decisions about demographically diverse examinees, we have developed an index, termed the *Consequential Validity Ratio* (CVR, Dumas et al., in revision). This index captures how well the scores from a given test can predict a criterion free from the undue influence of examinee demographics. Statistically, CVR is the ratio of the effect size of a focal measure (e.g., capacity) to the total variance explained by the capacity and participant demographics combined. The formula is written as:

$$CVR = \eta^2_{Test} \big/ R^2 \quad (5)$$

where $R^2$ is the total variance explained by the focal measure and demographics together, and $\eta^2$ is the effect size of the focal measure only. This ratio index takes the form of a decimal ranging from 0.00 to 1.00 and can be read as a proportion of the variance in the criterion prediction that is accounted for by the scores themselves (i.e., the signal) rather than the recruit demographics (i.e., the noise). In general, a higher CVR (i.e., a higher proportion of signal to noise) indicates better consequential validity of the focal measure.

***CVR Comparisons: DMM Capacity versus 1st ASVAB Score***

We calculated and compared CVRs from linear regression models predicting USCG competency with different focal measures (DMM capacity or the first ASVAB score) and the same set of demographic variables as predictors. All the calculated CVRs are summarized in Table 17 (see attached Validity Analyses Code and Output.pdf for effect sizes and calculation details). Given the very small percentages of recruits who took the ASVAB more than three times, this category was combined with the 3 attempts group in this analysis.

Table 17.
*CVR Comparisons*

| | CVR | |
|---|---|---|
| Number of ASVAB Attempts | First overall ASVAB score | DMM Capacity |
| 1 | 98% | 98% |
| 2 | 91% | 97% |
| 3 or more | 74% | 95% |

For the sub-sample that only took ASVAB once, $CVR_{1st\text{-}ASVAB}$ and $CVR_{capacity}$ were the same (98%), which indicates that there is not much incremental value of using capacity estimates in recruitment for this particular group of examinees. This was to be expected, and is because DMM is essentially redundant with the lone ASVAB score for recruits with one attempt given the limited information available with only one test score.

In order to determine if taking the ASVAB more than once allowed for fairer decisions to be made about demographically diverse recruits, we then calculated CVRs for the group of recruits that took the ASVAB exactly two times. As a result, the $CVR_{1st\text{-}ASVAB}$ was 91%, and $CVR_{capacity}$ was 97%. This indicates that, with only one extra ASVAB attempt (i.e., two attempts total) and DMM modeling, the consequential validity or fairness of the scores increased by nearly 6%.

Then we repeated the same analysis with the subset of recruits who took the ASVAB three or more times. For this group, the first ASVAB score had a CVR of 74%, indicating that demographics were a substantial contributor. In contrast, for the group that took the ASVAB three or more times, the CVR of the DMM capacity scores was 95%. This indicates that, when applying DMM modeling for recruits who take the ASVAB at least two extra times, the fairness of the assessment is increased by 21 percentage points (relative gain of nearly 30%).

### *What is the 'A School' Performance of USCG Recruits who Take ASVAB More Times?*

From the results in previous sections, we observed that examinees might improve their ASVAB score and enter a school via taking the test repeatedly. A major concern could be whether the recruits who have taken more ASVAB attempts to enter the training school would or would not perform as well as the others. If those recruits entering with multiple ASVAB attempts perform worse than their counterparts (i.e., recruits who only took the ASVAB once), the use of multi-timepoint measurement practice may have negative impacts on USCG recruitment and outcomes.

To check this potential issue, we fit a logistic regression model using the number of test attempts to predict a major training outcome, *Recycling Status*. Given that schools have their own specific requirements for recycling decisions, we controlled the effect of schools in the model. The number of test attempts was not a significant predictor of recycling status ($Z = -0.42$, $p = 0.68$, Odds Ratio = 0.97), which indicates having more ASVSB attempts was not associated with the training outcome of recycling. This implies that giving examinees multiple opportunities to take ASVAB would not decrease the quality of recruitment because they can perform as well as those who entered with taking the ASVAB once.

## Conclusions

Overall, the validity analyses indicated that the DMM Capacity score has good consequential validity, especially when comparing to the single-time point first ASVAB score. The descriptive findings in section B.1 have shown that the demographic minority groups may benefit most by taking the ASVAB multiple times. Based on the CVR comparisons, we further found the DMM capacity has the largest improvement of consequential validity for those who have taken the ASVAB repeatedly (e.g., relative gain of 30% for the subset of recruits who took the ASVAB three or more times). Taken together, allowing and encouraging examinees to take the ASVAB multiple times and utilizing DMM capacity scores can mutually support fair decisions and improve diversity in the USCG recruitment.

### Section B.5: Are Multiple ASVAB Attempts Better than Issuing Waivers?

As alluded to in Section A.3, one strategy to increase inclusivity of recruiting practices in the Coast Guard has been to lower minimum qualification composite scores or increase the frequency with which waivers are granted (particularly for recruits self-reporting as members of underrepresented groups). This would be juxtaposed against alternatives like recruits taking multiple ASVAB attempts until they can demonstrate that they can clear the standard.

In this section, we provide some evidence to address this question directly. In the Coast Guard data, 3,015 recruits had only one ASVAB attempt and were granted a waiver to attend their A school whereas 684 recruits had ASVAB multiple attempts, never received a waiver, and ultimately passed the standards for the A school they attended. This sets up a natural experiment such that recruits naturally self-selected into behavior associated with two alternative policies. Outcomes from these two naturally occurring groups can therefore be compared to discern whether the recruits with one ASVAB attempt and a waiver perform differently from recruits with multiple ASVAB attempts and no waiver.

The specific outcome of interest in this section is recycling, defined as recruits spending more time that prescribed in their training school. As a quick assessment to make sure that this definition accurately reflects recruits who required remedial time in their A School, we inspected the distribution of extra days in A School beyond the prescribed time for recruits classified as "recycling". 92% of recycled recruits spent at least 8 additional days in their A School, providing some evidence that recycling variable we calculated is capturing the need for remediation and not idiosyncrasies in the calendar (e.g., additional time is due to a federal holiday falling on a week day).

Table 18 shows the raw recycling rates for recruits with one ASVAB attempt and a waiver compared to recruits with multiple ASVAB attempts and no waiver. 45% of the recruits from the waiver group recycled, whereas only 35% of the non-waiver group recycled. The group differences in recycling rates were statistically significant ($G^2(1) = 22.95, p < .01, \phi = .08$), indicating the presence of an association between waiver group and recycling. These results showed that the candidates who entered schools through their own efforts (i.e., eventually meeting the A School standard through multiple ASVAB attempts) were less likely to recycle than those who received a waiver after only one ASVAB attempt.

Table 18.
*Waiver Group Differences in Recycling Rate*

| | | Multiple ASVAB, No Waiver | One ASVAB, Waiver | Total |
|---|---|---|---|---|
| Recycling | No | 445 (65%) | 1,661 (55%) | 2,106 (57%) |
| | Yes | 239 (35%) | 1,354 (45%) | 1,593 (43%) |
| Total | | 684 (19%) | 3,015 (81%) | 3,699 |

## Potential Confounders and Selection Bias

Table 18 presented the raw, marginal association between waiver group membership and recycling status, but the waiver groups formed naturally and were not randomly assigned. Therefore, there may be relevant selection effects that could confound proper interpretation of the association. One notable source is A School differences in recycling, which were previously discussed in Section B.4. For instance, the recruits attending the OS and ET schools recycled at much higher rates than did recruits attending the GM and DC schools. If the no waiver group systematically attended the schools with lower recycling rates like GM or DC, the decreased recycling rate in the no waiver group might be a selection artifact rather than a generalizable effect. That is, recruits from the no waiver group may recycle less because they happen to attend the schools with lower recycling rates rather than because of any tangible benefits of multiple ASVAB attempts.

We therefore tested allocation of recruits to A Schools with an $18 \times 2$ test of independence to determine whether there was an association between waiver group and A School or whether the groups where approximately balanced. The test was significant ($G^2(17) = 227.48, p < .001, \phi = 0.24$), indicating that A School and waiver group were significantly associated and that selection effects may be present such that recruits' desired A School may have affected their behavior and affected whether they choose to seek a waiver or attempt the ASVAB again. Additionally, previous analyses showed that certain types of recruits were more likely to select into multiple testing.

In the next section, we describe different methods that can account for possible selection effects in non-randomized data (Guo & Fraser, 2014) including regression adjustment (Cochran, 1968), inverse propensity score weighting (Lunceford & Davidian, 2004), propensity score matching (Rosenbaum & Rubin, 1985), and doubly robust augmented inverse propensity score weighting (Robins, Rotnitzky, & Zhao, 1994). We then apply consider which method is best and fit models to estimate recycling rate differences, adjusting for sources of self-selection.

## Accounting for Self-Selection into the Waiver and No-Waiver Groups

### Regression Adjustment

The classical approach to accommodating measured confounders in non-randomized data is to directly include the confounder as a covariate in the statistical model for the outcome. In the current context for modeling recycling, the model would be logistic regression with recycling as the outcome and waiver group as focal predictor and school attended as a covariate. The idea is that, by including the confounder as a covariate, the group effect will condition out any impact of the confounder and will be more representative of the true effect as if the data were randomized and all background variables were balanced between groups.

### Inverse Propensity Weighting

Propensity score analysis is a statistical method used to compare data from non-randomized groups as if the data could have been randomized (Austin & Stuart, 2015). This method is

popular for situations in which randomization is not feasible or ethical (e.g., people cannot ethically be assigned to a condition where they are forced to smoke cigarettes) or for observational studies where the groups form organically (Lipsey & Cordray, 2000).

The general idea of propensity score analysis using possible measured confounders is to create a model for self-selection into groups (Robins et al., 1994). That is, the model is a logistic regression with group status (waiver or no-waiver in our analysis) as the dependent variable. This is different from the regression adjustment method where confounders are included directly in the model for the outcome (i.e., recycling in our analysis) such that the model for the design and the model for the outcome are separated.

The propensity model yields a predicted probability that each recruit selected into the no-waiver group. This predicted probability is called the *propensity* and ranges from 0 to 1 where 0 means a person exhibited strong self-selection into the waiver group, 1 means a person exhibited strong self-selection into the no-waiver group, and 0.50 means there was relatively little self-selection and group status was essentially selected at random. After calculating a propensity for each recruit, the goal is to use the propensities to make the groups more comparable to a design that would have randomly assigned people to groups (Harder, Stuart, & Anthony, 2010).

In the inverse propensity weighting method, the confounders are not directly included in the model for recycling. (e.g., Williamson, Forbes, & White, 2013). Instead, the model for recycling *weights* each recruit by the inverse of their propensity such that recruits with strong evidence of self-selection are weighted less heavily and recruits with minimal evidence of self-selection are counted more heavily. The underlying idea is that recruits who chose a group mostly by chance are most valuable in the analysis because they represent behavior that would have been observed had group assignment been random. Conversely, recruits with strong evidence of self-selection are less valuable to the analysis because their group assignment was largely influenced by outside mechanisms, which is less consistent with random assignment. The main benefit of inverse propensity weighting is that all recruits are retained in the analysis, each recruit is just weighed different during model estimation.

**Propensity Score Matching**

Rather than using the propensities discussed in the preceding paragraph as weights, they can also be used to match similar people in different groups. The idea is that, if the goal is to adjust the data so that they look more like a randomized experiment, then creating two groups of recruits with similarly valued propensities should make the groups more directly comparable. With matching, this is achieved by discarding recruits without a suitable match in the other group such that recruits with strong evidence of self-selection are not included in the analysis because they are not representative of data that would be observed in a randomized experiment.

There is a large literature on different ways to match people based on their propensities (e.g., Abadie & Imbens, 2016; Ho et al., 2007). However, we only consider greedy matching in this analysis. Greedy matching is one-to-one, meaning that each recruit will only be matched to one other recruit and matches cannot be shared between multiple groups. The process starts with a randomly selected participant from the target group and the best match is selected from the

opposing group, even if the person in the opposing group would be a better match for a different person in the target group (this is what makes the method "greedy").

In greedy matching, the definition of "sufficiently close" is defined by the caliper (Rosenbaum & Rubin, 1985). The caliper is a user-defined value that determines the largest possible difference in propensity that can be used to still consider two recruits in opposing groups as matched. If there is no match in the opposing group within the specific caliper distance, then the recruit is dropped from the analysis. A common propensity score caliper is 0.25 standard deviations in the logit of the propensity (e.g., Rubin & Thomas, 2000), although more strict values like 0.10 can be selected in an attempt to increase the comparably of groups (Austin, 2011).

### Augmented Inverse Propensity Weighting

Regression adjustment includes possible confounders directly in the outcome model and propensity score weighting and matching include confounders in a separate model to estimate propensity. However, it is possible to include confounders in both models with augmented inverse propensity weighting (although both models need not contain the same set of confounders). This method is "doubly robust", meaning that for estimates to be correct, either the propensity model or the outcome model must be correctly specified, but not both (Bang & Robins, 2005). This is opposed to both regression adjustment (Glynn & Quinn, 2010) and inverse propensity weighting (Zhou, Matsouaka, & Thomas, 2020), which are both sensitive to potential misspecifications. The ability to be robust to misspecification is not without potential costs, however. The reduced bias associated with augmented inverse propensity weighting often comes at the cost of higher sampling variability, meaning that the method is susceptible to higher standard errors and lower power relative to previous methods that are properly specified (Kang & Schafer, 2007; Kurz, 2022).

### Comparing Recycling Between Waiver Groups, Accounting for Self-Selection into Groups

In our analysis, we use the augmented inverse propensity weighting approach discussed in the previous section. This method was selection because (a) the group sizes were unequal, so many common matching methods who cause the sample size to be reduced and (b) the data are administrative and we did not have access to interview recruits about their reasons for selecting into groups, so the double robustness property is especially advantages.

Possible confounders included in the propensity and outcome models largely overlapped and consisted of ASVAB score from the first attempt, self-reported non-White racial identity, self-reported Hispanic ethnicity, self-reported sex, and fixed effects for A school attended. Two- and three-way interactions between non-White identity, Hispanic identity, and sex were also included. A school was included because there are baseline differences in both the frequency of waivers and the recycling. Score from the first ASVAB attempt was included because self-selection in a waiver group may have been influenced by how far the recruit was from the entrance standard at a particular A school. The first ASVAB attempt was not included in the outcome model because admission is based on the final ASVAB score rather than the first, so the first score seems less relevant as a confounder of recycling. Both the propensity and the outcome model were binary logistic regressions.

The model was fit in PROC CAUSALTRT in SAS 9.4 and standard errors were estimated either with empirical methods (Stefanski & Boos, 2002) or bootstrapped with 1,000 replications. We also calculated confidence intervals with the normal-theory Wald method and with bias-corrected bootstrapping. The model treats confounders exogenously, so missing values are less straightforward to accommodate because they are conditioned out of the likelihood. Questions about race and ethnicity items were optional, so some recruits did not respond to these items (*N* = 677, 18.3%). As a sensitivity analysis, we also fit a version of the model that does not contain the non-White and Hispanic identity variables as predictors. All other predictors had no missing data and presented no issues.

Table 19 shows the estimated recycling rates for the no-waiver and waiver groups after accounting for possible confounders related to the outcome and self-selection into different groups. Table 19 is based on the 3,022 recruits that reported all demographic information. After accounting for possible confounders, the recycling rate in the no-waiver group was estimated to be 35.0% [95% Wald CI = 29.0% to 41.1%] compared to the 43.3% in the waiver group [95% Wald CI = 41.3% to 45.3%]. This difference of 8.3% is slightly smaller than the 10.0% difference observed in the raw percentages in Table 18, but is still statistically significant regardless of whether the difference is tested with empirical standard errors (Z = -2.64, p = .008), bootstrapped standard errors (Z = -3.02, p = .002), or the 95% bias-corrected bootstrapped confidence interval (interval = [-0.141, -0.029]).

Table 20 shows the results from the sensitivity analysis after removing confounders with missing values such that the analysis is able to include all 3,699 recruits. The difference in the recycling rate between waiver groups was smaller when removing non-White and Hispanic as confounders and the predicted recycling rate was 37.0% for the no-wavier group (95% Wald CI = [31.9%, 42.1%]) and 43.5% for the waiver group (95% Wald CI = [41.8%, 45.3%]). Nonetheless, the difference remained statistically significant if inference was conducted with empirical standard errors (Z = -2.45, p = .014), bootstrapped standard errors (Z = -2.73, p = .009), or the 95% bias-corrected bootstrapped confidence interval (interval = [-11.0, -1.3]).

This analysis provides consistent evidence that repeated ASVAB testing significantly reduces the rate of recycling relative to waivers among recruits that did not meet the A school standard after one ASVAB attempt. Given that there are significant costs associated with recycling, having recruits retake the ASVAB would appear to be more cost effective than granting waivers to recruits who do not meet entrance standards after one ASVAB attempt. There may also be potential psychosocial benefits to recruits if recycling were reduced (e.g., increased self-esteem and self-concept if the rate of failure were reduced).

Table 19.
*Recycling rates between recruits who received a waiver after one ASVAB attempt and recruits who retook the ASVAB until they met the standard, using augmented inverse propensity matching to adjust for sources of self-selection into groups. Analysis only includes recruits who reported all demographic information (N = 3,022)*

| | | Empirical | | | | Bootstrapped | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group | Recycling % | *SE* | 95% CI | *Z* | *p*-value | *SE* | 95% CI | *Z* | *p*-value |
| No Waiver | 35.0 | 3.1 | [29.0, 41.1] | | | 2.7 | [29.3, 40.4] | | |
| Waiver | 43.3 | 1.0 | [41.3, 45.3] | | | 1.0 | [41.3, 45.1] | | |
| Difference | -8.3 | 3.1 | [-14.4, -2.1] | -2.64 | .008 | 2.7 | [-14.1, -2.9] | -3.03 | .002 |

*Note:* The Difference row is calculated as (Waiver minus No-Waiver) such that negative numbers indicate that the no-waiver group had lower recycling rates. SE = Standard Error, CI = Confidence Interval

Table 20.
*Recycling rates between recruits who received a waiver after one ASVAB attempt and recruits who retook the ASVAB until they met the standard, using augmented inverse propensity matching to adjust for sources of self-selection into groups. Analysis all recruits, regardless of who demographic information reported (N = 3,699)*

| | | Empirical | | | | Bootstrapped | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group | Recycling % | *SE* | 95% CI | *Z* | *p*-value | *SE* | 95% CI | *Z* | *p*-value |
| No Wavier | 37.0 | 2.6 | [31.9, 42.1] | | | 2.3 | [32.7, 41.8] | | |
| Waiver | 43.5 | 0.9 | [41.8, 45.3] | | | 0.9 | [41.7, 45.3] | | |
| Difference | -6.5 | 2.7 | [-11.7, -1.3] | -2.45 | .014 | 2.4 | [-11.0, -1.3] | -2.73 | .006 |

*Note:* The Difference row is calculated as (Waiver minus No-Waiver) such that negative numbers indicate that the no-waiver group had lower recycling rates. SE = Standard Error, CI = Confidence Interval

### Waiver vs. No-Waiver Group Recycling Differences for Non-White Recruits

To address issues related to inclusivity specifically, we conducted the same analysis but restricted the analysis only to non-White recruits. This allows us to inspect differences in the recycling rate for non-White recruits specifically to determine if recycling rates remain discrepant in these demographic groups. This analysis focuses on the 681 recruits who either received a waiver after one ASVAB attempt or took the ASVAB multiple times without ever receiving a waiver *and* who self-reported a racial identity other than White. This subsample is 18% of the total sample in the previous section (21% if only counting recruits who reported any racial identity).

Table 21 shows the same comparison of recycling rates, specifically for non-White recruits. With no adjustment for school differences, non-White recruits who pursue multiple ASVAB attempts and never seek a waiver have a recycling rate 15.8 percentage points lower than non-White recruits who receive a waiver after one ASVAB attempt (32.2% vs. 48.0%, respectively), which is statistically significant and has a stronger effect size than the full sample ($G^2(1) = 10.01, p = .002, \phi = 0.12$).

Table 21.
*Waiver Group Differences in Recycling Rate for Recruits Reporting as Non-White*

| | | Multiple ASVAB, No Waiver | One ASVAB, Waiver | Total |
|---|---|---|---|---|
| Recycling | No | 293 (52%) | 80 (68%) | 373 (55%) |
| | Yes | 270 (48%) | 38 (32%) | 308 (45%) |
| Total | | 563 (83%) | 118 (17%) | 681 |

To account for sources of self-selection into the waiver groups, we applied the same model described in the previous subsection to this subsample of data. We fit the model two different ways: once using all demographic variables as potential confounders ($N$ = 608) and once without Hispanic as a potential confounder to include all 681 non-White recruits in the analysis. The results of the first model as shown in Table 22; the results of the second model are shown in Table 23.

In the model using Hispanic ethnicity as a potential confounder, after modeling sources of self-selection, the recycling rate in the no-waiver group was estimated to be 34.9% [95% Wald CI = 27.2% to 42.5%] compared to the 45.6% in the Waiver group [95% Wald CI = 41.4% to 49.8%]. This difference of 10.7% is smaller than the 15.8% difference observed in the raw percentages in Table 21, but is still statistically significant regardless of whether the difference is tested with empirical standard errors ($Z$ = -2.70, $p$ = .007), bootstrapped standard errors ($Z$ = -2.73, $p$ = .006), or the 95% bias-corrected bootstrapped confidence interval (interval = [-18.3, -1.7]).

In the model without Hispanic ethnicity as a potential confounder, the difference in the recycling rate between wavier groups was 9.6% and slightly smaller than the model that included Hispanic ethnicity as a possible confounder. The predicted recycling rate was 36.7% for the no-waiver group (95% Wald CI = [28.7%, 44.6%]) and 46.2% for waiver group (95% Wald CI = [42.3%, 50.2%]). The difference remained statistically significant if inference was conducted with empirical standard errors ($Z$ = -2.31, $p$ = .021), bootstrapped standard errors ($Z$ = -2.28, $p$ = .023), or the 95% bias-corrected bootstrapped confidence interval (interval = [-18.3, -1.6]).

This analysis provides evidence that the significant reduction in recycling rates remains specifically for recruits self-reporting a non-White identity. Therefore, repeated ASVAB testing without waivers appears to better address diversity initiatives by reducing recycling among members of underrepresented racial identities, in addition to cost effectiveness advantages mentioned in the previous section. Similar analyses were also performed for female recruits and Hispanic recruits, but no significant differences in recycling were found.

Table 22.
*Recycling rates between recruits who received a waiver after one ASVAB attempt and recruits who retook the ASVAB until they met the standard, using augmented inverse propensity matching to adjust for sources of self-selection into groups. Analysis includes non-White recruits who responded to the Hispanic ethnicity question (N = 608)*

| | | Empirical | | | | Bootstrapped | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group | Recycling % | *SE* | 95% CI | *Z* | *p*-value | *SE* | 95% CI | *Z* | *p*-value |
| No Waiver | 34.9 | 3.9 | [27.2, 42.5] | | | 4.2 | [27.7, 43.8] | | |
| Waiver | 45.6 | 2.1 | [41.4, 49.8] | | | 2.1 | [41.3, 49.7] | | |
| Difference | -10.7 | 4.0 | [-18.5, -2.9] | -2.70 | .007 | 4.3 | [-18.3, -1.7] | -2.73 | .006 |

*Note:* The Difference row is calculated as (Waiver minus No-Waiver) such that negative numbers indicate that the no-waiver group had lower recycling rates. SE = Standard Error, CI = Confidence Interval

Table 23.
*Recycling rates between recruits who received a waiver after one ASVAB attempt and recruits who retook the ASVAB until they met the standard, using augmented inverse propensity matching to adjust for sources of self-selection into groups. Analysis only includes all non-White recruits (N = 681)*

| | | Empirical | | | | Bootstrapped | | | |
|---|---|---|---|---|---|---|---|---|---|
| Pathway | Recycling % | *SE* | 95% CI | *Z* | *p*-value | *SE* | 95% CI | *Z* | *p*-value |
| No Waiver | 36.7 | 4.1 | [28.7, 44.6] | | | 4.0 | [28.4, 44.6] | | |
| Waiver | 46.2 | 2.0 | [42.3, 50.2] | | | 2.1 | [42.3, 50.3] | | |
| Difference | -9.6 | 4.2 | [-17.7, -1.5] | -2.31 | .021 | 4.3 | [-18.3, -1.5] | -2.28 | .023 |

*Note:* The Difference row is calculated as (Waiver minus No-Waiver) such that negative numbers indicate that the no-waiver group had lower recycling rates. SE = Standard Error, CI = Confidence Interval

# Final Report Chapter C: Discussion and Recommendations for Future Work

## Section C.1: Methodological Advances Accomplished in this Project

Before detailing the specific ways in which this current research project illuminated patterns among coast guard recruits, we will first detail some of the methodological innovations that were accomplished in this project that allowed us to observe meaningful patterns in these data. It is important to note that essentially none of the methods applied in this project were ready 'out of the box'. Instead, each aspect of the current work was tailored specifically for the project at hand.

### Data Merging and Organization

Before this project began, the USCG did not necessarily keep a full dataset of all ASVAB scores for a recruit, linked to the A school that a recruit attended, as well as their outcomes during that training. Because all of these aspects of the recruitment, selection, and training processes were relevant to our project, we worked with the USCG and other service branches to assemble and organize such a dataset. As one of the key deliverables produced during the earlier stages of this project, we provided this organized and labeled dataset back to the USCG along with a data dictionary that was designed to support the work of future analysts.

### Latent Coast Guard Competency Scoring

One key reason why USCG members take the ASVAB is for selection into training programs called 'A Schools'. Each of the A schools has their own set of standards for admission that pertain to the overall ASVAB or a combination of ASVAB sub-scales. What this means is that some A schools are easier or harder to get into than others, and some recruits are offered admissions to more A schools than others. So, we conceptualized this situation as a classically utilized psychometric model (called a Rasch model; Wright & Stone, 1979) that simultaneously modeled the difficulty for admission at each of the A schools and each recruits' latent competency to be admitted. Each of these two parameters were placed on the same scale, so they can be directly compared.

We see this innovation as being particularly useful for future work in understanding the way that ASVAB standards either intentionally or unintentionally rank both the A schools in terms of difficulty and the recruits in terms of competency. Such a model could be used to investigate the effects of shifting (or hypothetically shifting) the ASVAB standard for a particular school, and opens the door to a straightforward understanding of the way A school recruitment is influenced by ASVAB scores.

### Modeling Self-Selection to Multiple ASVAB Testing Occasions

One of the most unique and complex aspects of the data we were working with here was the situation in which USCG members self-selected to take the ASVAB only once (which about 90% of recruits chose to do) or to take the ASVAB multiple times (the remaining 10%). This self-selection process is actually not entirely uncommon in educational testing and is present in other higher-stakes assessment contexts such as college admissions tests. However, it is only

very rarely dealt with in the statistical literature related to education (see Matta & Soland, 2019 for an example with English language testing).

This situation meant that over the course of the ASVAB administrations, the test-scores were missing for a theoretically important reason: the recruit had chosen not to continue to test either because they were satisfied with their score or because they did not believe they could improve. In a sense, we were therefore able to think of the individuals who did not continue to take the ASVAB as 'dropping out' of the data set before they needed to, and the individuals who took the ASVAB more times as persisting through that drop-out process. We saw a metaphor between this situation and survival processes that occur in biostatistics (such as when a study participant dies before the study is complete) and therefore we drew on methods that are typically more associated with the biosciences (i.e., survival modeling; Ohno-Machado, 2001).

So, when we built the DMM that formed the central part of this research, we built it to simultaneously model the growth in test scores that was exhibited by those recruits who took the test multiple times as well as the dropout process by which the recruits exited the dataset and therefore could no longer improve their test scores. More specifically, we configured the model both with a known-class mixture model that described which recruits took the test only once and which took the test multiple times. Then, among the recruits who took the test multiple times, we used a survival model to describe how many testing occasions they chose to take before dropping out of the testing program. It is important to note that this statistical innovation had never been accomplished in the context of modeling learning capacity and dynamic measurement, and it formed an important part of why the current project was able to be successful.

**Conceptualizing Differences in Both 'A School' Selection Standards and Recruits' Goals**

In other repeated-measures educational testing contexts, there might be a set criterion score that respondents must reach in order to achieve competency and stop taking the test. In these contexts, statistical modeling can be accomplished more easily and readily because the operant question really becomes how many attempts does an individual need until they reach competency? However, the current research took place in a much more complex and nuanced context, because each of the recruits entered the testing program with different goals and expectations related to the training they wanted to pursue, and each training program (i.e., A school) had differing standards for admission. In addition, it was also possible for recruits to be admitted to a program without actually meeting the ASVAB standard (i.e., receive a waiver), and it was possible for recruits to choose to pursue a training program that was not the most difficult they were admitted to (i.e., over 200 people in the current dataset chose to be an Yeoman despite testing highly enough to be admitted to Marine Science Technician training).

Because of this inherent complexity of the dataset, we were not able to use some existing modeling solutions for repeated-measures data. One such solution is called a time-to-criterion model (Johnson & Hancock, 2019), and it takes a specific interest in how many attempts, or how much time, it takes an individual to reach a pre-determined level of competency. Instead, we formulated our more nuanced model in the tradition of DMM where we modeled an individual-specific learning capacity score that was informed both by their persistence in the testing program (i.e., how many times they took the test), how high their raw test scores were, and how

much they improved in between each testing occasion. This model configuration allowed us to meet the needs of the USCG while also adequately handling the high level of complexity in these data.

## Section C.2: What did the DMM Show about USCG Recruits?

In this section, we highlight several key findings from this study, and draw on the detailed Results presented above in Chapter B to justify the points we make here.

### Minoritized Recruits Persisted through more ASVAB Attempts

One unique and interesting aspect of the data that we analyzed here is that—unlike most longitudinal or repeated-measures data we typically work with (e.g., ECLS-K, Tourangeau et al., 2009)—the USCG recruits had the choice to take the ASVAB only once, or to take it multiple times. This situation created a kind of natural experiment in which most of the recruits (~90%) only had one ASVAB attempt to analyze, while a much smaller group (~10%) had repeated-measures data. For this reason, we carefully modeled the phenomenon of persistence vs. drop-out in the DMM we fit here, with the intention of identifying the individuals who were opted-in to more testing attempts. As was presented in detail in Chapter B, we uncovered a significant effect of demographics on persistence through multiple ASVAB attempts: with individuals with more historically minoritized identities within the USCG (i.e., females, non-White recruits) choosing to take the ASVAB more times.

Understanding this effect is complex and requires some conjecture on our part, especially because recruits are likely to choose to take the ASVAB multiple times for varying reasons. For instance, a recruit who scores relatively low on their first attempt might choose to take the test again, while another recruit who scored well, but who held particularly high standards for themselves, might also choose to re-take the test. Conversely, recruits may choose to cease taking the ASVAB for a variety of reasons including satisfaction with their score, discouragement, or practical constraints on how many times they could re-take the test. For these reasons, it is not straightforward to say definitively why this phenomenon occurred in the USCG dataset. For our part, we would hypothesize that historically minoritized candidates for the USCG may have had fewer educational experiences in their past that specifically prepared them for the ASVAB, making their first attempt at the test largely a process of familiarization with the format and content. Then, these minoritized candidates for the USCG must have generally had an inclination that their first attempt was not their best work and felt motivated to persist with more testing attempts. In our view, this level of persistence represented a critically important source of resilience for these individuals, and their persistence may also be a meaningful resource for the USCG to tap into in the future.

### Recruits Typically Improved as they Re-Tested

Despite the inherently longitudinal paradigm of the current research, it is important to note that the ASVAB is actually designed to be taken only one time by each recruit. Typically within the field of psychometrics, when a test is designed to be a higher-stakes single-timepoint measure, it is also thought that scores are unlikely to improve as the result of re-testing. In fact, the correlation between test and re-test scores has classically been used as evidence of the reliability of a test for many decades (DeVellis & Thorpe, 2021). In a Dynamic Measurement paradigm, this assumption of test-retest rank-order preservation is not made (Dumas et al., 2020), and therefore our current work here was much more flexible than the more classic perspective.

What we found was that, over their ASVAB attempts, recruits typically improved substantially. Our model generated a learning rate score for every individual recruit, and for the vast majority of individuals (i.e., 99.5%) that rate was positive, indicating that they improved over time. This pattern of growth was required for us to model learning capacity using the DMM, but it also indicates a fundamentally crucial finding: the recruits who choose to re-take the ASVAB are not doing so for no reason. They are re-taking the ASVAB and scoring higher than they did during their previous attempt. We further found that female recruits ($M$ =5.14, $SD$ =1.03) had higher learning rate scores than male ($M$ = 5.01, $SD$ = .98) on average; the non-White group ($M$ = 5.18, $SD$ = .99) had higher mean learning rate than the White group ($M$ = 4.96, $SD$ = .98); and Hispanic recruits ($M$ = 5.16, $SD$ = 1.10) had higher mean learning rate than non-Hispanic ($M$ = 4.97, $SD$ = .98). These results consistently show that the minority groups can improve more over time, which indicates that encouraging candidates to take the ASVAB multiple times may resulted in a more diverse body of recruits for the USCG.

Precisely why and how recruits were able to improve across their ASVAB attempts is not totally clear. We would hypothesize that the growth is driven in large part by recruits initially familiarizing themselves with the ASVAB: learning the format of the test and becoming aware of the content. Of course, there are other explanations as well. Perhaps some recruits engaged in effortful learning experiences (e.g., studying) between ASVAB attempts in order to improve their score. Another possibility is that some recruits experienced profound stress and anxiety during their initial attempt, and as they took the test more times, that anxiety lessened and allowed them to perform better. Still, a combination of these factors may be at work for some individuals. Only a more targeted study that followed-up with individuals who improved steeply during their ASVAB testing period would be able to know this for sure.

**More ASVAB Attempts Did Not Predict Recycling in 'A' school**

One legitimate concern that inevitably arises in issues of re-testing for recruitment and selection is whether or not the individuals who earned their spot in their training program as a result of re-testing perform as well as their peers who earned their spot with only a single test attempt. For this reason, our team tested this question in the USCG data and found that those recruits who earned their spot in 'A' school as a result of re-testing did not have any higher probability of needing to recycle in that program. 'Recycling' is a term used by the USCG to indicate that a recruit did not adequately meet the standards of their training in their first attempt, and they required additional time in the training program in order to meet the standards.

This finding is important because it demonstrates that allowing recruits to re-take the ASVAB should not be thought of as a policy that potentially lowers the quality of recruit selection into training programs. It also suggests that, as recruits re-take the ASVAB, their scores not only improve but also become better indicators of their learning capacity, therefore actually aiding the USCG in selecting individuals into training programs validly. So, far from being a source of error in the recruitment and selection process, it appears that re-takes of the ASVAB are providing a crucially important signal to the USCG about who is qualified for what training programs.

**Learning Capacity was Estimable from ASVAB Data**

Before we began this project, modeling the learning capacity of individuals who opted-in to re-tests, rather than being required to test a certain number of times, had never been attempted. In addition, our previous work with DMM has typically utilized datasets that contain substantially more repeated-measures for each individual (usually five or more time-points; e.g., Dumas & McNeish, 2017). For this reason, the current research was conceptualized as a feasibility study, or a proof-of-concept, in order to determine whether a dataset like this one from the USCG could be used to model learning individuals' learning capacity. As has been closely demonstrated in Chapter B above, we found that learning capacity was indeed estimable in the current data.

Our success in modeling the learning capacity of USCG recruits was driven by a number of statistical innovations we developed over the course of this project (detailed in section C.1) including the simultaneous modeling of the persistence/drop-out and growth processes that were present in the dataset. Given these innovations, we were able to answer a core question associated with this research in the affirmative: Yes, learning capacity can be modeled using DMM in this dataset of USCG recruits' ASVAB scores.

**Learning Capacity was a Fair and Valid Predictor of USCG Competency**

After building the DMM and estimating the learning capacity parameters, we were able to generate learning capacity scores for every individual recruit in the dataset. With that scoring complete, the next question logically becomes: Are learning capacity scores valid? Especially given the current focus on diversity and inclusion within the USCG, we paid special attention to the fairness or consequential validity of the learning capacity scores. Our findings were very encouraging and suggested an important level of fairness and validity to the learning capacity scores generated here.

In order to ascertain the fairness of the learning capacity scores, we created an index that we named the *Consequential Validity Ratio* (CVR; Dumas et al., in revision). This quantity was designed to capture the degree to which a psychometric score (such as learning capacity scores, or a more typical single-timepoint test score) is capable of predicting an important outcome variable, free from the undue influence of demographics such as gender or race. Using this index, we were able to demonstrate that, for participants who chose to take the ASVAB three or more times, the predictive link between their first time-point score and their USCG competency was relatively highly affected by their demographic background. In contrast, for the same group of participants, their DMM-estimated learning capacity score was a better predictor of the USCG competency outcome, and that predictive pathway was much less affected by demographic background. In terms of relative percentage points, DMM learning capacity scores were able to improve upon the fairness of recruits' initial ASVAB attempt by 30%.

The reasons why this phenomenon of improved fairness was observed are clear to us, given the theoretical underpinning of DMM. We would posit that, as recruits choose to take the ASVAB more times, important information is gathered about those recruits (e.g., their new ASVAB scores, and their rate of improvement between attempts). When this important information is

added to a psychometric model, it becomes possible to estimate much more meaningful quantities (i.e., learning capacity) about recruits. To put this another way, we would argue that the first ASVAB attempt is a relatively weak indicator of recruits' actual potential to learn and succeed in the USCG, because that first attempt is affected by many sources of error that obscure individuals' true potential (e.g., the unfamiliarity of the testing format; systematic differences in past learning experiences across demographic groups). As participants continue to take the ASVAB over time, more and more information is gained about their true potential, and a more fair and valid estimate of their learning capacity becomes possible with DMM.

## Section C.3: How Should Current USCG Recruitment Practice be Changed?

Here, we draw on the key findings presented above in Section C.2, to suggest some key amendments to current USCG recruitment practice using the ASVAB.

### More Encouragement for Recruits to Re-Take the ASVAB

All of the results of this current study suggest that recruits choosing to re-take the ASVAB is a positive thing. More time-points to the ASVAB allow for recruits' improvement trajectories to be observed, and their true learning potential to be inferred based on those trajectories of improvement. So, we strongly believe that re-testing on the ASVAB is not a source of error or noise in the recruitment process, instead it is a source of important information or signal about each recruit who elects to re-test.

### Waivers Should Not be Given After Only One ASVAB Attempt

In section B.5 above, we presented a detailed analysis where we demonstrated that the USCG recruits who received a waiver after only a single time of attempting the ASVAB were consistently and substantially more likely to not pass their training program on the first try, and be required to re-take their training (i.e., they recycled). In comparison, the group of recruits who never received a waiver, but earned their way into their A school by re-taking the ASVAB, were consistently less likely to be recycled. What this pattern indicates is that giving waivers to recruits after their first ASVAB attempt appears to lower the quality of USCG recruitment and increase the likelihood of failure in A school.

These differences were particularly pronounced for recruits reporting a racial identity other than White. This indicates that increasing the frequency of waivers with the intent to increase inclusivity will not necessarily be effective or in the best interest of the recruit or the USCG because it results in more recycling. Increased recycling can consume more resources for the USCG (time and money, in particular) and may decrease feelings of cohesiveness and increase perceptions of tokenism if recruits feel they were admitted based on their demographic characteristics rather than their ability (e.g., Perez & Strizhko, 2018).

One potentially problematic aspect of the waiver-based system currently in use at the USCG is that, based on our analysis, it may be somewhat arbitrarily applied. That is, it was impossible to determine solely from the data *why* any individual recruit had been granted a waiver and another individual recruit had not. In many cases, recruits with the same raw ASVAB point difference from the standard they needed to attend their eventual A schools had different solutions applied to them: some received waivers and others re-took the ASVAB. Still others may have chosen to attend an A school with a lower entrance standard. We have no way of knowing why certain recruits were selected to receive a waiver and why others were encouraged to re-take the test. What we do know is that re-taking the test, rather than receiving the waiver after the first ASVAB attempt, was associated with better outcomes in A school, especially for non-White recruits.

One aspect of this empirical pattern that is important to understand is that different A schools had different baseline recycling rates. But, in our analysis, we statistically controlled for this potential confound using a variety of modern analytic methods, and in every case the pattern remained consistent. Those recruits who received a waiver after only one timepoint were more likely to recycle, even after carefully controlling for possible confounders related to self-selection. For this reason, we feel we can confidently make the recommendation to the USCG to suspend the practice of waiving ASVAB scores after only one attempt. Instead, our analysis suggests that recruits should be encouraged to re-take the test in order to earn their place in A school.

**More Standardized Learning Experiences between ASVAB Attempts**

As of now, it was not known to our team what methods recruits may have utilized to support their learning and score improvement in between their ASVAB attempts. We were able to clearly demonstrate that recruits did indeed tend to improve across re-takes of the test, but we cannot know for certain how that improvement was accomplished. In the future, we would suggest the provision of relatively standardized learning materials to recruits after their initial attempt of the ASVAB in order to maximize the chances that their score will improve during their re-take. The additional standardization of learning materials or experiences between ASVAB attempts would have a number of other benefits as well, including the possible alleviation of demographic differences (if certain groups have greater or lesser access to learning materials or tutors). In general, a shifted focus on recruitment for learning capacity would benefit from additional attention paid to the materials and experiences that allow recruits to learn, especially if their score is low at first.

**Conceptualize Re-Taking the ASVAB as Persistence**

In our experience as an authorship team, we have found that re-taking any higher-stakes exam including the ASVAB, sometimes carries a negative stigma for recruitment and selection. For this reason, some recruits (or recruiters) might conceptualize re-taking the test as a form of failure or a sign of intrinsically lower learning capacity. However, our results suggest that this is not the case. Instead, we would conceptualize the re-taking of the ASVAB as an important indicator of persistence and motivation to succeed in the USCG: an indicator that should be valued in and of itself. We would hypothesize that re-taking the ASVAB was framed differently as ‘Persisting on the ASVAB,’ and many more recruits would elect to take the test multiple times. For our part, we highly value individuals’ grit and determination in achieving their goals, and we find that opting-in to a repeated measures testing program rather than dropping out after only one test attempt to be a highly desirable form of persistence.

## Section C.4: Future Directions for Research with USCG Data

Despite the major strides made during this project in understanding the psychometric patterns in USCG recruitment and selection data, a number of key future directions exist that our team might closely investigate within USCG data in the future. Here, we highlight three next-steps in this line of inquiry that we foresee to be particularly fruitful.

### Person-Centered Modeling Approaches

The DMM model utilized in this current study is fundamentally *variable-centered*, meaning that it posited key latent variables that are hypothesized to cause the patterns in the dataset. In this case, those key latent variables were the slope, or improvement rate between ASVAB testing occasions, and the capacity, or predicted upper limit on an individual's ASVAB scores. Of course, other statistical paradigms exist for elucidating psychologically-relevant aspects of these data, and one key paradigm that may be worth examining is a *person-centered* approach.

In person-centered approaches to modeling, the underlying causation of the patterns in the dataset is not hypothesized to be driven by latent variables that participants have in their minds (e.g., learning capacity), but rather driven by latent sub-classes of individuals that participants belong to (Dumas et al., 2021). So, person-centered approaches share many similarities with more typical variable-centered latent variable models but discretize the latent variable as a way to identify latent subgroups of participants. For instance, many educators in a variety of contexts conceptualize their students as belonging to different types whose educational strengths, weaknesses, and needs differ (Forsten et al., 2002). A person-centered modeling approach is therefore specifically designed to detect these underlying subgroups of students and clarify the true differences between them.

In light of the current goals of the USCG, a person-centered modeling approach might be a fruitful next-step, because understanding the membership of particular recruits within latent sub-classes could provide useful information for educational decision-making. In addition, the current focus on diversity and inclusion within the USCG may be well-served by a person-centered modeling approach. Interestingly, latent sub-classes of recruits may or may not correlate strongly with more obvious demographic differences (e.g., race or gender), but how those demographic differences influence latent class membership may speak volumes about who USCG recruits are, and what their educational needs are.

### Item-level ASVAB Differential Item Functioning Analysis

The current focus on diversity, equity, and inclusion in the recruitment and selection processes of the USCG implies that the USCG might be concerned about potential bias of the ASVAB. The DMM applied in the current study is designed to detect and describe recruits' learning capacity—and we know from past work that learning capacity tends to be a much more fair indicator of individual potential to learn than individual test scores (Dumas & McNeish, 2017)—but it was not designed to detect bias on individual items of the ASVAB test.

In order to conduct a full study of possible bias on the ASVAB test, our team would require item-level data from every test respondent (i.e., we would need to know which items on the test each recruit got right or wrong), and those detailed data were not available to our team for the current study. In the future, however, it would be feasible for our team to examine every individual item on the ASVAB across all salient demographic groups and determine whether or how demographic variables influenced the probability of a correct answer, while controlling for overall ability level. In psychometrics, this type of analysis is called *differential item functioning* (DIF; Osterlind & Everson, 2009) and is very popular in our research community and among educational testing firms. DIF is a necessary but not sufficient condition for bias, and DIF research is an important step in determining if an item is or is not biased.

Because our team did not have access to the item-level ASVAB data to conduct a full DIF analysis, the DMM we fit here assumed that each individual ASVAB test score was not systematically biased. Instead, our model was designed to mitigate fairness and validity issues caused by certain participants having had less educational opportunity in the past, which is a pervasive social issue in the US, but different than bias on the test itself (Dumas & McNeish, 2017). It is important to note that our results here are meaningfully interpretable regardless of what an item-level DIF study of the ASVAB might reveal, but we do recommend an item-level DIF analysis as a next-step in this line of work.

**Interpretation of Growth Norms**

The DMM we built and applied in the current study was designed to quantify how much individual recruits improved between ASVAB testing occasions, and what their learning capacity (i.e., the upper limit on growth) would be. So, the DMM was able to produce scores for every individual recruit that represents those two latent variables: each recruit in the dataset now has a learning rate and a learning capacity score from the DMM. Using these scores, we are able to show which recruits have a faster learning rate than others, and we can also show which recruits have higher or lower learning capacity in comparison to others. However, one interesting future direction that this current study did not encompass is providing interpretable norms so that practitioners (i.e., USCG recruitment and selection personnel) might interpret a recruits' learning rate or learning capacity.

Within the field of psychometrics, producing clear and interpretable norms is an important way that researchers such as us serve the practitioners who utilize our findings. In fact, there are areas of work within education or psychology where professionals spend most of their time interpreting test scores in light of norms: norms that were previously developed by psychometricians. For this reason, our field thinks deeply about how to produce, present, and explain normative information about the scores we estimate to practitioners (DeVellis & Thorpe, 2021). In the future, it would be interesting and potentially important to create clearly interpretable norms from the current USCG DMM, and train recruitment personnel on how to interpret them.

This possible future direction around the interpretation and use of DMM norms is much more than a research endeavor, it is the only way that higher-stakes decisions about recruitment and selection can be made by practitioners on the basis of learning capacity. In the future, we

envision a potential world where USCG recruitment is conducted with learning capacity carefully considered for every individual. The results of this study suggest that such a future would be more fair and valid than the current practice (e.g., issuing waivers), and would concomitantly support diversity and inclusion efforts. However, that future is only possible if recruitment and selection personnel are trained to interpret normative scores from a DMM. For this reason, we suggest the creation of interpretable norms for learning capacity, and the training of USCG personnel to utilize those norms for recruitment and selection, as a critical next step.

**Acknowledgements**

This research was made possible through the Testing Modernization Funds, Office of the Under Secretary of Defense for Personnel and Readiness/M&RA/MPP(AP). The authors wish to express appreciation to Dr. Sofiya Velgach, (Assistant Director of Testing Standards, OUSD), and Dr. Cyrus K. Foroughi (Research Scientist, Naval Research Laboratory), for funding support. We also wish to express special thanks to the Coast Guard's CDR Morgan T. Holden, and Brett F. Ayer (Office of Strategic Workforce Planning and HR Analytics), and LCDR Blake Leedy (Force Readiness Command [FC-Tms]) for access/assistance with the Coast Guard dataset.

## References


Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, *84*(2), 781-807.

Albert, P. S., & Follmann, D. A. (2000). Modeling repeated count data subject to informative dropout. *Biometrics*, *56*(3), 667-677.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. AERA.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399-424.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661-3679.

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods*, *16*(2), 149-165.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962-973.

Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research*, *43*(2), 210-236.

Bayroff, A. G., & Fuchs, E. F. (1970). *The armed services vocational aptitude battery* (Vol. 1161). US Army Behavior and Systems Research Laboratory.

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods*, *9*(1), 30-52.

Blozis, S. A., & Cho, Y. I. (2008). Coding and centering of time in latent curve models in the presence of interindividual time heterogeneity. *Structural Equation Modeling*, *15*(3), 413-433.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Britannica (1993). *Encyclopaedia Britannica*. Britannica.

Calkins, A., & Asch, B. J. (2022). *What Happened to Military Recruiting and Retention of Enlisted Personnel in 2020 During the COVID-19 Pandemic?*. RAND CORP.

Canz, T., Piesche, N., Dallinger, S., & Jonkmann, K. (2021). Test-language effects in bilingual education: Evidence from CLIL classes in Germany. *Learning and Instruction*, *75*, 101499.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24* (2), 295-313.

Colford, M., & Sugarman, A. J. (2016). Millennials and the military. *Hoover Institution*.

Connell, C. (2011, Oct 22). *ASVAB score requirements changing for "A" schools.* https://www.mycg.uscg.mil/News/Article/2816942/asvab-score-requirements-changing-for-a-schools/

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. *Psychological Methods*, *7*(1), 41-63.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C, 43*(1), 49-73.

Dong, Y., Dumas, D., Clements, D. H., & Sarama, J. (2022). Developing a Trajectory Deviance Index for Dynamic Measurement Modeling. *Journal of Experimental Education*. Advance online publication. https://doi.org/10.1080/00220973.2022.2044280

Dumas, D. G., & McNeish, D. M. (2017). Dynamic measurement modeling: Using nonlinear growth models to estimate student learning capacity. *Educational Researcher, 46*(6), 284-292.

Dumas, D. G., Dong, Y., & Leveling, M. (2021). The zone of proximal creativity: What dynamic assessment of divergent thinking reveals about students' latent class membership. *Contemporary Educational Psychology, 67*, 102013.

Dumas, D., Dong, Y., & McNeish, D**.** (in revision). How fair is my test?: A ratio statistic to help represent consequential validity. *European Journal of Psychological Assessment.*

Dumas, D., McNeish, D. & Greene, J. A. (2020). Dynamic measurement: A theoretical-psychometric paradigm for modern educational psychology, *Educational Psychologist, 55*(2), 88-105, DOI: 10.1080/00461520.2020.1744150

Dumas, D., McNeish, D., Sarama, J., & Clements, D. (2019). Preschool mathematics intervention can significantly improve student learning trajectories through elementary school. *AERA Open, 5*(4), 233285841987944. https://doi.org/10.1177/2332858419879446

Dumas, D., McNeish, D., Schreiber-Gregory, D., Durning, S. J., & Torre, D. M. (2019). Dynamic measurement in health professions education: rationale, application, and possibilities. *Academic Medicine, 94*(9), 1323–1328. https://doi.org/10.1097/ACM.0000000000002729

Elliott, J. (2003). Dynamic assessment in educational settings: Realising potential. *Educational Review, 55*(1), 15-32.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods, 16*(1), 1-16.

Feng, Y., Hancock, G. R., & Harring, J. R. (2019). Latent growth models with floors, ceilings, and random knots. *Multivariate Behavioral Research, 54*(5), 751-770.

Feuerstein, R. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques*. University Park Press.

Flores, Alfinio. (2007). Examining disparities in mathematics education: achievement gap or opportunity gap? *The High School Journal, 91*(1), 29–42. https://doi.org/10.1353/hsj.2007.0022

Forsten, C., Grant, J., & Hollas, B. (2002). *Differentiated instruction: Different strategies for different learners*. Crystal Springs Books.

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis, 18*(1), 36-56.

Grigorenko, E. L. (2009). Dynamic assessment and response to intervention: Two sides of one coin. *Journal of Learning Disabilities, 42*(2), 111-132.

Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in M plus and SAS. *Structural Equation Modeling, 16*(4), 676-701.

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications.

Guo, S., & Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and applications* (2nd edition). SAGE publications.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in crossnational surveys of educational achievement* (pp. 58-79). National Academies Press.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. Wiley.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234-249.

Harring, J. R., Strazzeri, M. M., & Blozis, S. A. (2021). Piecewise latent growth models: beyond modeling linear-linear processes. *Behavior Research Methods, 53*(2), 593-608.

Haywood, H. C., & Miller, M. B. (2003). Dynamic assessment of adults with traumatic brain injuries. *Journal of Cognitive Education and Psychology, 3*(2), 137-163.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199-236.

Huber, P. J. (1973). Robust Regression: Asymptotics, conjectures, and Monte Carlo. *Annals of Statistics 1* (5), 799–821.

Jensen, A. R. (1985). Armed services vocational aptitude battery. *Measurement and Evaluation in Counseling and Development, 18*(1), 32-37.

Johnson, T. L., & Hancock, G. R. (2019). Time to criterion latent growth models. *Psychological Methods, 24*(6), 690-707.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22*(4), 523-539.

Kapp, L. (2002). *Recruiting and Retention in the Active Component Military: Are There Problems?* Congressional Research Service.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine, 17*(23), 2723-2732.

Kim, S. Y., Mun, E. Y., & Smith, S. (2014). Using mixture models with known class membership to address incomplete covariance structures in multiple-group growth models. *British Journal of Mathematical and Statistical Psychology, 67*(1), 94-116.

Kuhfeld, M., Gershoff, E., & Paschall, K. (2018). The development of racial/ethnic and socioeconomic achievement gaps during the school years. *Journal of Applied Developmental Psychology, 57*, 62–73. https://doi.org/10.1016/j.appdev.2018.07.001

Kurz, C. F. (2022). Augmented inverse probability weighting and the double robustness property. *Medical Decision Making, 42* (2), 156-167.

Lipsey, M. W., & Cordray, D. S. (2000). Evaluation methods for social intervention. *Annual Review of Psychology, 51*(1), 345-375.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*(431), 1112-1121.

Lim, K., Hall, K. C., Keller, K. M., Schulker, D., Mariano, L. T., Matthews, M., Saum-Manning, L., Hill, D, Crosby, B., Payne, L.A., Cottrell, L., & Aranibar, C.A., (2021). *Improving the Representation of Women and Racial/Ethnic Minorities Among U.S. Coast Guard Active-*

*Duty Members.* Homeland Security Operational Analysis Center, FFRDC by RAND Corporation.

Liu, F., & Eugenio, E. C. (2018). A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. *Statistical Methods in Medical Research, 27*(4), 1024-1044.

Long, J., & Ryoo, J. (2010). Using fractional polynomials to model non-linear trends in longitudinal data. *British Journal of Mathematical and Statistical Psychology, 63* (1), 177-203.

Lord, J., & Mayer, R. (2020, Dec. 02). *CO/OIC ASVAB Waiver Authority for "A" School*. Presentation to the Personnel Readiness Task Force.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine, 23*(19), 2937-2960.

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research, 32*(3), 215-253.

Matta, T. H., & Soland, J. (2019). Predicting time to reclassification for English learners: A joint modeling approach. *Journal of Educational and Behavioral Statistics, 44*(1), 78-102.

McNeish, D. & Dumas, D. (2017). Nonlinear growth models as measurement models: A second-order growth curve model for measuring potential, *Multivariate Behavioral Research, 52*(1), 61-85, DOI: 10.1080/00273171.2016.1253451

McNeish, D., & Dumas, D. (2021). A seasonal dynamic measurement model for summer learning loss. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 184*(2), 616-642.

McNeish, D., & Matta, T. (2018). Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Behavior Research Methods, 50*(4), 1398-1414.

McNeish, D., Bauer, D. J., Dumas, D., Clements, D. H., Cohen, J. R., Lin, W., ... & Sheridan, M. A. (2021). Modeling individual differences in the timing of change onset and offset. *Psychological Methods*, advance online publication.

McNeish, D., Dumas, D. G., & Grimm, K. J. (2020). Estimating new quantities from longitudinal test scores to improve forecasts of future performance. *Multivariate Behavioral Research, 55*(6), 894–816. https://doi.org/10.1080/00273171.2019.1691484

McNeish, D., Dumas, D., Torre, D., & Rice, N. (2022). Modelling time to maximum competency in medical student progress tests, *Journal of the Royal Statistical Society, Series A*, advance online publication.

McNeish, D., Harring, J.R., & Dumas, D. (under review). A multilevel structured latent curve model for disaggregating student and school contributions to learning. *Statistical Methods & Applications.*

Mehta, P. D., & Neale, M. C. (2005). People are variables too: multilevel structural equations modeling. *Psychological Methods, 10*(3), 259-284.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5*(1), 23-43.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). ACE & NCME.

Muthén, B., & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics, 30*(1), 27-58.

Muthén, L.K. and Muthén, B.O. (1998-2022). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Ohno-Machado, L. (2001). Modeling medical prognosis: survival analysis techniques. *Journal of Biomedical Informatics*, *34*(6), 428-439.

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609-1623.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Sage Publications.

Preacher, K. J., & Hancock, G. R. (2015). Meaningful aspects of change as novel random coefficients: A general method for reparameterizing longitudinal models. *Psychological Methods*, *20*(1), 84-101.

Quester, G. H. (2005). Demographic trends and military recruitment: Surprising possibilities. *The US Army War College Quarterly: Parameters*, *35*(1), 11.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846-866.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.

Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics*, *41* (1), 103-116.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*(450), 573-585.

Stafford, D. E., & Griffis, H. S. (2008). A review of millennial generation characteristics and military workforce implications. *Center for Naval Analysis, http://www.cna.org/documents D*, *18211*.

Sterba, S. K. (2014). Fitting nonlinear latent growth curve models with individually varying time points. *Structural Equation Modeling*, *21*(4), 630-647.

Sterba, S. K. (2017). Partially nested designs in psychotherapy trials: A review of modeling developments. *Psychotherapy Research*, *27*(4), 425-436.

Sterba, S. K., Preacher, K. J., Forehand, R., Hardcastle, E. J., Cole, D. A., & Compas, B. E. (2014). Structural equation modeling approaches for analyzing partially nested data. *Multivariate Behavioral Research*, *49*(2), 93-118.

Thiébaut, R., Jacqmin-Gadda, H., Babiker, A., Commenges, D., & Cascade Collaboration. (2005). Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Statistics in Medicine*, *24*(1), 65-82.

Thorndike, E. L. (1919). A standardized group examination of intelligence independent of language. *Journal of Applied Psychology, 3*(1), 13–32. https://doi.org/10.1037/h0070037

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K): Combined user's manual for the ECLS-K eighth-grade and k–8 full sample data files and electronic codebooks* (NCES 2009-004). National Center for Education Statistics.

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, *14* (3), 809-834.

United States Coast Guard (2021, Oct. 22). *Changes to Armed Services Vocational Aptitude Battery (ASVAB) Requirements for "A" School Attendance* https://www.dcms.uscg.mil/ppc/news/Article/2819883/changes-to-armed-services-vocational-aptitude-battery-asvab-requirements-for-a/

US Department of Defense (1999). *Computing IRT reliabilities for the ASVAB Student Testing Program*. Defense Manpower Data Center.

Welsh, J. R., Kucinkas, S. K., & Curran, L. T. (1990). *Armed Services Vocational Battery (ASVAB): Integrative review of validity studies* (Technical Report No. 90-22). Air Force Systems Command.

Williamson, E. J., Forbes, A., & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine*, *33*(5), 721-737.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement*. Mesa Press.

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44* (1), 175-188.

Yao, L. (2022, Mar. 8). *Preliminary Findings for the ASVAB Enlistment Testing Program (ETP): Potential Impact of COVID-19*. Presentation to the MAPWG.

Yin, P. (2022, Mar. 8). *Preliminary Findings for the ASVAB Career Exploration Program (CEP): Potential Impact of COVID-19*. Presentation to the MAPWG.

Zhou, Y., Matsouaka, R. A., & Thomas, L. (2020). Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, *29*(12), 3721-3756.