NORC at the University of Chicago

CHRR AT THE OHIO STATE UNIVERSITY

# TASK 2.4.2
# DEPARTMENT OF DEFENSE CONTENT PANEL REPORT FINAL

September 8, 2022

## Contract Deliverable

**Presented by:** Roxanne Wallace, PMP, NORC

**Authored by:** Judy Hellerstein (chair), Joseph Altonji, Andrew Ho, Joe Rodgers, Paul Sackett, Dan Segall, Michael Walker

# Document Change Log

| Published Date | Document Version No. | Pages Affected | Description of Revision | Author |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

# 1. Introduction

The National Longitudinal Surveys (NLS) are a significant, long-running program of the United States (U.S.) Bureau of Labor Statistics (BLS), designed to support research into how Americans navigate changes in the economy and transition through various life course stages. As the youngest NLS cohort members are now entering their 40s, the BLS seeks to begin a new cohort of adolescents, targeted for fielding in 2026. This NLSY26 cohort will enable researchers to understand new trends in labor market experiences, education, and a wealth of other factors that are affecting this new generation.

BLS contracted with NORC at the University of Chicago and CHRR at The Ohio State University on an NLSY Needs Assessment to provide BLS with topical content and methodological inputs that a future design team can use to create an NLSY26 survey responsive to key research goals. As part of this Needs Assessment, NORC convened a content panel on Department of Defense (DoD) initiatives, comprised of federal and non-federal subject matter experts, to provide BLS with high-level recommendations about leveraging potential synergies between BLS and DoD in the collection of measures of cognition, personality, other abilities (e.g. mechanical), and career interests of youth. The content panel met multiple times between May and July 2022 to discuss recommendations and tradeoffs around content and survey design for BLS to consider for the new cohort.

The rest of this report is organized as follows: Section 2 and Section 3 describe the panel's review of issues related to the administration, collection, and potential uses of DoD assessments for a new NLSY26 cohort and, concurrently, for two other samples of interest to DoD drawn simultaneously with a new NLSY26 cohort. Section 2 describes the current and potential future DoD assessments along with the benefits to BLS, researchers, and DoD from administration of these tests to a new NLSY26 cohort and two other related DoD samples. The three key assessments discussed are: (1) Armed Services Vocational Aptitude Battery (ASVAB); (2) Tailored Adaptive Personality Assessment System (TAPAS); and (3) Find Your Interests (FYI), a measure based on Holland's RIASEC. Section 3 separately describes the DoD and BLS perspectives on various issues related to the NLSY26 cohort and the household sample from which it would be drawn, including (1) two possible supplemental samples for DoD and the benefits of the administration of DoD assessments to these samples and to the NLSY26 sample; (2) possible DoD assessments for consideration to various samples; (3) the possible ages of the NLSY26 sample and the timing and frequency of DoD assessments to the cohort; (5) test proctoring and potential accommodations; (6) incentivizing participation and score reporting; and (7) data sharing between DoD and BLS. Section 4 discusses specific panel recommendations related to (1) the ages of samples to be drawn for DoD purposes and related timing of assessment administration; (2) the age range of the new NLSY26 cohort; (3) the recommended assessments to be administered to the NLSY26 cohort; and (4) the age of first administration to the NLSY26 cohort and the frequency of (re)administration.

# 2. Topic-Related Recommendations for the New Cohort

## Emerging research themes, social trends and policy changes that are relevant for the content area

The Department of Defense sponsors three types of assessments for qualifying applicants into the military and for high school students participating in the Career Exploration Program (CEP). These assessments measure three broad areas of individual differences: (a) Cognitive ability and other skills as measured by the ASVAB, (b) Personality (measured by TAPAS), and (c) Vocational Interests (measured by the FYI). These three sets of measures are used by the DoD to qualify applicants into the military, assign qualified applicants into specific military jobs, and to help guide career exploration for high school students participating in the ASVAB Career Exploration Program.

**ASVAB**

The Armed Services Vocational Aptitude Battery (ASVAB) consists of 10 tests spanning four domains: Verbal, Math, Science and Technical, and Spatial (Table 1). Four of the tests (AR, WK, PC, and MK) are combined into an Armed Forces Qualification Test (AFQT) composite that is used to qualify applicants for military service. There are two modes of administration for the ASVAB: a computerized adaptive testing (CAT) version, and a paper-and-pencil (P&P) version. Nearly all military applicants take the CAT-ASVAB, while the P&P-ASVAB is offered to most students participating in the Career Exploration Program[1]. Each of the 10 CAT-ASVAB tests consists of 10 or 15 adaptively selected items and on average the entire battery takes about 90 minutes to complete. To help maximize testing efficiency and precision, test items are adaptively chosen for each test-taker based on their test performance. Consequently, the CAT-ASVAB can achieve the same or higher levels of precision as compared to the P&P-ASVAB with about 40% fewer test items.

EXHIBIT 1. ASVAB Description[2]

| Test | Description | Domain |
|---|---|---|
| 1. General Science (GS) | Knowledge of physical and biological sciences | Science/Technical |
| 2. Arithmetic Reasoning (AR) | Ability to solve arithmetic word problems | Math |
| 3. Word Knowledge (WK) | Ability to select the correct meaning of a word presented in context and to identify best synonym for a given word | Verbal |

[1] There is a desire by DoD over time to increase the use of CAT-ASVAB and reduce the reliance on P&P-ASVAB in the High School Career Exploration Program.

[2] Taken from ASVAB Fact Sheet (www.officialASVAB.com)

| | | |
|---|---|---|
| 4. Paragraph Comprehension (PC) | Ability to obtain information from written passages | Verbal |
| 5. Mathematics Knowledge (MK) | Knowledge of high school mathematics principles | Math |
| 6. Electronics Information (EI) | Knowledge of electricity and electronics | Science/Technical |
| 7. Auto Information (AI) | Knowledge of automobile technology | Science/Technical |
| 8. Shop Information (SI) | Knowledge of tools and shop terminology and practice | Science/Technical |
| 9. Mechanical Comprehension (MC) | Knowledge of mechanical and physical principles | Science/Technical |
| 10. Assembling Objects (AO) | Ability to determine how an object will look when its parts are put together | |

The CAT-ASVAB is utilized in two different modes: proctored and unproctored. The proctored CAT-ASVAB is offered with separately timed test sections and is administered at testing locations managed and proctored by DoD-sponsored personnel. The unproctored CAT-ASVAB can be taken at a time and location of the test-taker's choosing and has no explicit time limit, other than the requirement that it must be completed within a three-day window after starting[3].

Scores on the ASVAB subtests used for enlistment qualification are reported on scales derived from data collected in conjunction with the NLSY97. This data collection, based on an 18-23 year old group, is referred to as the Profile of American Youth (PAY97). ASVAB test scores are reported on a scale with a mean of 50 and SD of 10 (relative to the PAY97 group). Composite scores are also reported on scales derived from PAY97, where some composite scales report percentile scores (relative to the PAY97) and other scales are set to have a mean of 100 and SD of 20 for the target population represented by the PAY97 group. The scale for Coding Speed was set in the same manner since it was part of the ASVAB when data were collected for PAY97.

ASVAB scores for students participating in the Career Exploration Program are reported on a different set of scales derived from the PAY97 10th, 11th, and 12th grade cohorts[4]. Note that these cohorts are distinct from the 18-23 year old group used to norm the Enlistment ASVAB, and were based on youth expected to be enrolled in grades 10, 11, or 12 as of the fall of 1997. For each test and composite, nine sets of grade-specific and gender norms were developed: three each for 10th, 11th, and 12th grade (as well as for Males, Females, and combined sex groups within each grade).

Potential additions to the ASVAB are under consideration at the DoD; these are discussed on page 13.

[3] In order to be used for Military enlistment, scores from the unproctored version must be verified by a short proctored verification test. Because the verification test is not used by DoD independent of the unproctored exam, the panel did not consider it as an ASVAB substitute for the NLSY26 sample.

[4] See ASVAB Norms for the Career Exploration Program

**Personality**

The Department of Defense also administers the Tailored Adaptive Personality Assessment System (TAPAS) to select military applicants for use in selection and classification. The battery measures up to 27 facets of personality and includes measures of military-specific traits. Many of these traits are subsumed within the Big 5 Factor domain. TAPAS is computer delivered and utilizes a multidimensional pairwise preference model to reduce the influence of faking. Test-takers are presented with statement pairs matched on endorsement rates and social desirability and are asked to select the statement that most closely represents them. The assessment is also adaptive, so statement pairs are selected to maximize measurement efficiency (achieve precise measurement on several dimensions with fewer items than would be required by a fixed conventional approach). Each military Service uses a different version of TAPAS, although these versions have overlapping sets of facets. There is an effort underway to consolidate the different versions into a single standardized instrument (with a common set of facets) that will be administered to all applicants across all Services in the future.

The facet scores for TAPAS are based on an underlying IRT scale where the origin and unit are aligned to the mean and SD of an applicant calibration group. Composite score scales are also reported on scales derived from military applicant groups.

**Career Interests**

The Department of Defense offers the Find Your Interest (FYI) inventory to high school students as part of the Career Exploration Program. As a measure of career and vocational interests, the FYI is designed to help students learn about their career-related interests and is based on Holland's RIASEC theory of career choice. The FYI assesses an individual's resemblance to each of the six RIASEC (Realistic, Investigative, Artistic, Social, Enterprising, and Conventional) types described by Holland (1997). The assessment contains 90 items, with most students completing it within about 15 minutes, and provides both gender-based and gender-combined percentile scores for interpretation. Because the interest inventory was developed after the NLSY97/PAY97 efforts had been conducted, the norms for FYI were developed from a different group and are based on a nationally representative sample of schools.

## Selected topics considered for data collection with the new cohort

**For the new NLSY26 cohort**

A new NLSY26 cohort will consist of youth who are either exclusively or almost exclusively below the age of 18. As such, DoD assessments provided to these respondents could include the ASVAB and FYI, as DoD currently administers them to 10$^{th}$-12$^{th}$ graders as part of CEP. A personality assessment that is age appropriate would round out comparability with the three main military assessments, but TAPAS has not previously been administered to youth in this age group and has multiple dimensions that are military-specific.

The ASVAB and earlier versions of FYI were also administered to previous NLSY cohorts, allowing for potential cross cohort studies. As the NLSY26 respondents age, repeat administration of the ASVAB and a personality assessment could be considered.

**Additional DoD samples**

When the household sample is drawn for the NLSY26, two additional samples of older youth/young adults could be simultaneously sampled for DoD purposes, as was done previously for the PAY97 group (see Section 3 for more information). Specifically:

1. A sample of 18-23 year-olds, individuals of military-enlistment age. This sample would be in the age range in which DoD currently administers the ASVAB and TAPAS to enlistees.

2. A sample of 10th-12th graders (or youth in the age range of those ages). DoD administers the ASVAB and FYI to youth in this group. Depending on the exact ages of the NLSY26, there could be overlap between some of NLSY26 cohort and this sample, and NLSY26 respondents could be included in this sample.

## Related foundational data important for studying labor market outcomes

Our panel was focused on DoD-related assessments. There is a well-established literature that details the relationship between labor market outcomes and the AFQT score (see Aughinbaugh, Pierret & Rothstein, 2015 for a discussion of some key research). Documenting other related foundational data was considered out of scope of the panel.

## Related foundational data important for studying other later life outcomes

Our panel was focused on DoD-related assessments. Documenting other related foundational data was considered out of scope of the panel.

## Key areas of disparities and inequalities that should be measurable

The panel recognized from the outset that there is a voluminous body of impactful research demonstrating the importance of the AFQT as a measure of cognitive skills in the NLSY surveys and in assessing how skills develop during childhood, how dimensions of abilities and traits affect educational choices and

outcomes, and how they affect adult outcomes—most notably multiple dimensions of labor market outcomes, but also outcomes such as criminal behavior, adult health, and family formation. The AFQT has often have been used in research examining outcomes for white men, but also, importantly, to understand race, sex, and ethnic differences in outcomes. A full literature review would be prohibitively time- and space-consuming, and a literature sampling likely would not do justice to the breadth and depth of these studies. More generally, the importance of measuring cognitive ability, non-cognitive ability, and personality traits, along with a discussion of some key research, are addressed in the report of the K-12 content panel. This panel therefore began with a shared baseline understanding of the continued importance of the collection of the AFQT as part of the ASVAB battery in the new NLSY26 cohort. The panel also made note of relatively more recent research emphasizing the importance of non-AFQT dimensions of the ASVAB in the determination of labor market outcomes (see, e.g. Prada & Urzúa, 2017, Speer, 2017, and Light & McGee, 2015) because the non-AFQT subtests allow for the assessment of a wider variety of skills.

The panel also noted the importance of the measurement of non-cognitive skills and personality traits in the NLSY in the understanding of outcomes. Again, the panel did not find it necessary to discuss this at length or provide motivation via citations.

Perhaps least well known is the importance for research purposes of obtaining information about work preferences, at least with respect to the NLSY. A key question is the role that preferences across occupations and between market work and family play in shaping educational and labor market choices, as well as marriage and fertility choices. For example, Arcidiacono (2004) and Zafar (2013) find that gender differences in preferences and expectations play a key role in gender differences in college major choice. It is for this reason that the panel also recommends that the FYI be administered to the NLSY26. The panel also recommends that, if feasible, the results of the administration of the Interest Finder to the NLSY97 be released by DoD to BLS and incorporated into the NLSY97 data so that they can be used for research purposes, including possibly cross-cohort comparisons.

# 3. Survey Design-Related Recommendations for the New Cohort

## Degree of inclusion of recommended topics in NLSY79 and NLSY97

Because BLS (including stakeholders served by the NLSY datasets such as the research community) and DoD have somewhat different goals, we address each perspective separately below. It is important to recognize that the final recommendations listed in section 4 are structured to meet the collective needs of the BLS (and its stakeholders) and DoD. We begin with the DoD perspective:

## DoD Perspective

### Background

The Department of Defense has partnered with BLS in both previous NLSY79 and NLSY97 efforts. Because the population of interest was largely different and non-overlapping with BLS efforts, the data collected for DoD purposes has been routinely labeled as PAY80 and PAY97. PAY80 data were collected during the NLSY79 effort, while PAY97 data were collected as a part of the NLSY97 effort.

The PAY97 group, consisting of data from 10th, 11th, and 12th grade students and from 18-23 year old respondents, was used to:

a. *Provide updated norms for the ASVAB*. These norms provided the capability to translate ASVAB test scores and composite scores to percentile equivalents relative to contemporary enlistment (18-23 year old) and student (10th-12th grade) populations.

b. *Inform military recruiting policies and goals*. Knowing the distribution of ASVAB scores in the 18-23 year-old population (for the whole population and by race and geographic region) provides military workforce planners valuable data. Information regarding the qualified reservoir of youth available to fill current and future roles can inform both military recruiting and force structure policies and practices.

c. *Assess trends across time*. Work based on data collected in both the PAY80 and PAY97 studies has been conducted by DoD, including work focusing on how shifting demographics over this period might have influenced changes in the demographics of qualified military youth.

d. *Revise the ASVAB score scales (both subtest and composites)*. The PAY97 data were used to revise both subtest and composite ASVAB score scales. ASVAB test scales were constructed so that the scaled standard scores (on the 1997 metric) have a mean of 50 and a SD of 10 for the 18-23 year-old population. Composite score scales (including AFQT and service composites) were also updated using PAY97 data, with some composites (like the AFQT) assigning updated percentile scores and others deriving T-score-like transformations with means and SDs constrained to have means of 100 and SDs of 20 in the PAY97 youth population.

   In 2004, the ASVAB PAY97 scales replaced the PAY80 scales. This score-scale revision required large amounts of stakeholder discussion, coordination, impact assessments, and linking analyses (between new and old score scales), and, in general, created a great deal of consternation. Even though this linking could have been used to adjust cut-scores and benchmarks, DoD decided to implement the new score-scale without adjusting AFQT score ranges used to define key quality benchmarks that drove recruiting goals. As a result, some applicants qualifying for service or enlistment bonuses on the old scale were no longer qualified on the new scale. This score-scale shift had the effect of raising enlistment standards, but without a corresponding observable shift in test scores. Because of a relatively good recruiting

environment during this period and because of the relatively small shift in score-scale, DoD was able to absorb the negative recruiting impact.

In previous NLSY data collection efforts, DoD administered a form of ASVAB designed to be parallel to existing operational forms. Because of concerns over context effects, particular features of the test such as time limits, test-lengths, proctoring, and testing mode (P&P versus computer) were specified to match the version of the ASVAB that was in operational use during this period.

**Future Assessments**

DoD's greatest need moving forward is for two samples from which national norms can be developed: a) a sample of 18-23 year old for the full set of tests making up the ASVAB and for the TAPAS personality measure, and b) a sample of 16-18 year old ($10^{th}$–$12^{th}$ graders) for ASVAB and for the Find Your Interests (FYI) occupational interest measure. ASVAB and TAPAS are currently used for military entry, and thus updated norms for ASVAB and an initial set of national norms for TAPAS are needed to support decision making. ASVAB is also used, along with the FYI tool, in DoD's high school Career Exploration Program, and norms are needed for these measures for these populations.

Since TAPAS facet scores are influenced by which facets are included in the battery, the norming would benefit from a version of TAPAS that was used jointly by all Services. There are efforts currently underway to develop a joint-Service TAPAS battery that will contain a set of facets used by all Services. This version would be a good candidate for inclusion in the NLSY2026.

FYI scores on each of the RIASEC are translated to percentile scores based on a nationally representative sample of high schools. Students taking the FYI are encouraged to focus on the top scores (interest areas) based on these percentile scores to explore jobs with similar RIASEC emphases. For many of the same reasons ASVAB could benefit from a renorming (i.e., datedness of current norms), DoD has an interest in developing new norms for FYI based on the NLSY2026.

DoD also has expressed a strong interest in repeat administration of the ASVAB in age ranges that delineate according to military career stage: pre-enlistment ($10^{th}$-$12^{th}$ grade); enlistment age (18-23); and potentially somewhat later since the ASVAB is used for specialty reclassification. Repeat administration will improve DoD's understanding of ASVAB score stability over time.

## BLS Perspective

From the perspective of the BLS and users of the NLSY datasets, the goal of including measures of individual differences (including cognitive measures, social skills, and other constructs such as personality and vocational interests) is to support high quality research about the predictors and causes of labor market (and other life course) outcomes and potentially the effect of labor market and other life course outcomes on these skill measures.

With regards to age of first administration of an assessment, the panel noted tensions among competing objectives:

- Administer all tests at a young age (12-16) to help avoid possible attrition that can come from waiting until later rounds and to allow more direct comparisons with the NLSY97 cohort. Earlier assessments also help avoid ceiling effects.
- Test later (10th–12th grade) since there can be rank shifting in cognitive abilities between older and younger respondents which could obscure estimates of the importance of abilities if based on testing at a younger age. In addition, ASVAB, TAPAS, and FYI were developed and validated for use with older youth and young adults. There is limited evidence of which the panel is aware[5] of the validity of these instruments for assessing cognitive abilities, personality, and vocational interests for the younger age range, and (obviously) no evidence for the cohort that would constitute the NLSY26.

With regards to the selection of assessments and their value for predicting and understanding labor market outcomes, the panel made the following observations:

- For norming purposes, DoD must administer the entire ASVAB battery, whereas BLS could consider administering only select tests contained within the battery such as the AFQT tests (AR, WK, PC, and MK). However, while AFQT scores are clearly most important for research purposes for cognitive ability, other tests contained in the ASVAB have also proven to be important predictors of labor market outcomes as discussed above.
- TAPAS facet statements were designed for use on older youth and young adults. The validity of scores on TAPAS for the younger group has not been systematically studied.
- TAPAS measures a large number of personality dimensions that are specifically targeted to the needs of DoD, but it also includes traits that are aligned with Big 5 personality dimensions. To date, TAPAS has not been administered to a non-military sample.
- Research on vocational interests both within and across the NLSY97 and the NLSY2026 could be facilitated by the analysis of Interest Finder data collected as part of the NLSY97 study and the release of FYI data for the NLSY26 cohort.
- The FYI provides a measure of vocational interests grounded in the widely used RIASEC framework, useful for studying the connections between pre-employment interests and post-employment career choice as well as the fertility and gender-based differences discussed on page 8.
- Repeat administration of assessments as a cohort ages allows for the study of the stability of cognitive ability and personality measures over the lifespan.

There is significant interest among researchers in understanding how occupation and career interests and expectations during childhood and early adulthood map into realized educational and labor market outcomes. Recent studies in the United States often focus on populations studied in college (see e.g. Arcidiacono et al., 2012; Arcidiacono et al., 2020; Wiswall and Zafar, 2021), a narrow focus that is

[5] Dahlke et al., (2018) for the SAT and McBride et al., (2000) for early administration of the ASVAB both report high – but not perfect – degrees of test stability.

largely due to data limitations. There also is older research that examines dimensions of occupational aspirations of youth in the NLSY79 (and the original cohorts) and outcomes (e.g. Levine and Zimmerman, 1995; Levine and Rubinstein, 2017) as well as asking about determinants of those aspirations (e.g. Hoffman, 1987). Results from RIASEC assessments have been found to predict adult labor market outcomes in other countries (e.g., Stoll et al., 2017), and research on the stability of RIASEC scores over the life course is ongoing (Hoff et al., 2018). Thus, we expect that FYI scores would support an expansion of this type of research in the United States.

## Methodological issues to consider on recommended topics

The panel discussed a number of methodological issues relevant to both the BLS and DoD objectives.

**Proctoring**

In order for ASVAB scores to be used for enlistment purposes, DoD requires that either: (a) the full ASVAB be administered under in-person proctored conditions, or (b) scores from an unproctored version of the ASVAB be verified by a short, proctored verification test. TAPAS is also administered in a proctored setting. DoD representatives reported that higher ASVAB scores are obtained in unproctored settings for a key group of test takers (Segall, 2016) and have concluded unproctored testing is untenable for purposes of determining national norms. Therefore, the panel recommends that the ASVAB and TAPAS be administered in proctored settings only. For the enlistment age sample, BLS will have to align with DoD protocols, but for the NLSY26 cohort BLS may be able to consider remote proctoring options.

There are two options regarding proctoring: live in-person proctoring and remote proctoring by computer video. We find that we cannot offer an evidence-based recommendation regarding the comparability of live vs. remote proctoring for ASVAB and TAPAS. While there is a sizable literature comparing proctored vs. unproctored testing, the literature comparing live versus remote proctoring is limited. A critical feature of studies we can locate (e.g. Andreou et al., 2021; Cherry et al., 2021; Kim & Walker, 2021; Weiner & Hurtz, 2017) is a lack of random assignment to live vs. remote proctoring conditions. The studies involve operational high stakes testing programs in the domains of professional licensure or educational admission. They either compare one cohort tested under one form of proctoring with a later cohort tested under the other, or compare settings where, due to differential availability, live proctoring was used for some examinees and remote for others. Thus, selection bias is a plausible alternative explanation for differences across proctoring conditions. Kim and Walker (2021) offer the most sophisticated methodology used to date, using techniques that match samples on external variables (e.g., demographics) to better equate groups. Note also that those studies are in high-stakes settings where examinees have a strong incentive to obtain high scores, in contrast to the low-stakes setting for participation in NLSY, and generalizability of findings would be an issue even if the sample selection issues were solved. The panel agreed that NLSY field interviewer proctoring could be as an option for the NLSY26 sample. However, the panel had concerns about the length of the field interview if a field interviewer proctored exam is part of the process.

We suggest resolution of this issue prior to a decision about testing mode for a new NLSY sample. We do not claim to have conducted an exhaustive search for literature on this topic; perhaps there is useful work of which we are unaware. We believe that a randomized experiment comparing live and remote proctoring with a sample comparable to that intended for the NLSY would be reasonably straightforward. Kim and Walker (2021) were pessimistic about the prospects of randomization, but they were writing about doing so in an operational testing environment where examinees might resist random assignment to testing environments. Such a study would be much more straightforward in a low-stakes environment.

**Accommodations for Testing**

The Department of Defense does not provide accommodations for applicants testing for enlistment qualification but does allow for some accommodations for those testing in the High School Career Exploration Program (such as more time and reading assistance following 504 Plans). Accommodations should be made for the administration of the ASVAB to the NLSY26 respondents and to a 10th-12th grade high school DoD-specific sample to improve respondent access, in accordance with the accommodations granted to the students by their schools. A variable indicating that a test was accommodated should be recorded in the NLSY sample so that if re-administration occurs at a later date with potential changes in the availability of accommodations, this can be taken into account. Accommodations should not be given to a DoD-specific sample of 18-23 year old. National norms can thus be developed by DoD consistent with existing accommodation policies.

**Incentivizing Participation**

In a pretest prior to the NLSY97/PAY97, two incentives/performance bonus studies (McBride et al., 2000) were conducted to evaluate the influence of incentives to go to a Sylvan Test Center and bonuses based on test performance. In the Participation Incentives/Performance Bonus (PIPB) study, the authors examined the effect of differential bonuses ($25, $60, $75) on participation. They reported that participation rate increased directly with the amount of the monetary incentive, but the rate (62%) for the highest payment of $75 still fell short of the 90% participation target.

In a follow-up experiment, the authors studied the effect of randomly assigning respondents to receive larger differential incentives of $75 and $100 and concluded:

> Nothing in the participation rate or cost data of this experiment suggested any practical benefit to offering more than $75 as an incentive to participate in the psychometric testing planned for the 1997 Profile of American Youth (PAY97). The $75 incentive was as effective as $100 in terms of participation rate and was slightly less costly overall. Therefore, $75 was recommended as the PAY97 incentive amount. (p. 26)

The panel recommends the review of past relevant studies of participation incentivization along with current OMB policies involving the size and caps of respondent payments. Available information from

past studies of surveys and PAY97 might provide sufficient information to guide the specification of appropriate payments.

**Score Reporting (to Respondents)**

According to the Standards for Educational and Psychological Testing (2014), only scores that are used to make decisions need to be reported to the test-taker. However, BLS and DoD might want to consider providing scores and related interpretative materials to respondents to help foster and maintain high participation rates. If a decision is made to provide scores to test takers, two reporting options exist: (a) report scores on existing scales using existing norms, or (b) report scores on possibly updated score scales using new norms developed with NLSY2026 data. With regards to Option (a), BLS and DoD might consider granting NLSY2026 participants access to the ASVAB Career Exploration Program website (www.ASVABprogram.com). This site provides a description of the Career Exploration Program as well as detailed score interpretation materials for ASVAB and FYI scores. While updated normative information offered by Option (b) can provide additional useful interpretive information to participants, this normative information will require analyses of the full set of NLSY2026 data. Consequently, there could be a significant lag between the time of testing and the reporting of scores to individual respondents. For this reason, the panel recommends providing percentile score reports relative to the NLSY97 cohort (Option a) in addition to, or instead of, reporting scores relative to the NLSY2026 cohort (Option b).

**Data Sharing**

A data sharing agreement between the Department of Defense and BLS should be developed and signed well in advance of data collection, considering the extensive lead-time typically required for developing and approving these agreements. Possible topics for inclusion in the agreement include:

- Data Security Safeguards
- Compliance with DoD Systems of Records Notice (SORN). Assuming the assessments will be delivered online using DoD servers and infrastructure, the SORN should be reviewed and updated if necessary to ensure that NLSY2026 data can be appropriately collected and housed in existing databases currently approved for military applicants and high school students.
- Data Fields Provided by DoD. The agreement should specify which fields will be transmitted to BLS by DoD (e.g., test scores, composite scores, item responses, item latencies, etc.). DoD might not want to transmit some testing information such as item responses, keys, and item text due to the possibility of compromise. Historically, the same forms used in the NLSY studies have been reused in separate studies, so it is useful to restrict the dissemination of particular item information to help safeguard test security.
- AFQT Composite Scores. AFQT scores (as calculated and used by DoD) for NLSY2026 respondents should also be included.
- Scores for the ASVAB and AFQT. If the ASVAB changes, we recommend DoD conduct an equating study and provide actual ASVAB and AFQT scores and linked scores that adjust for changes in the test.

## Relevant alternative data sources to capture recommended topics

The Department of Defense has several additional cognitive tests that are administered to select applicants or are under consideration for inclusion in future versions of the ASVAB. These include:

- *Coding Speed*. A test of speeded ability administered to Navy applicants.
- *Cyber*. A test of information technology knowledge, including areas of networking and telecommunications, computer operations, security and compliance, and software programming and web design.
- *Complex/Abstract Reasoning*. A test of non-verbal abstract reasoning where respondents are asked to identify the missing element in a pattern.
- *Mental Counters*. A test of working memory.

There may be other assessments under consideration as well. The panel could not make recommendations about specific tests that are still under development, but these tests may be useful for future research and the panel suggests that BLS should continue to monitor new testing initiatives in coordination with DoD. DoD should work closely with BLS to consider the relevance and timing of the fielding of a new NLSY26 cohort as pertains to possible data collection of assessment outcomes for the national samples drawn as part of the new cohort.

# 4. Top Ranked Topic- and Survey Design-Related Recommendations

## Prioritized recommendations

The panel recommends the administration of the ASVAB and FYI to youth (under the age of 19), including the NLSY26, and the ASVAB and TAPAS to DoD-specific samples. Exhibit 2 summarizes the key recommendations of the DOD panel with regard to timing of assessment administration, with additional details provided below.

EXHIBIT 2

| | ASVAB | TAPAS | TIPI | FYI | NLSY97 comparison |
|---|---|---|---|---|---|
| Age 18-23 | 2026 | 2026 | | | ETP |
| 10th-12th grade | 2026 | | | 2026 | STP |

| | | | | | |
|---|---|---|---|---|---|
| Main NLSY26 cohort | 2026 (older members); 2028-29 (younger members); re-administered before age 24, around age 30, and around age 40 (perhaps AFQT only) | | 2026 and with any ASVAB re-administration | Concurrent with initial ASVAB | ASVAB and Interest Finder: Full sample in round 1 (ages 12-17)<br>TIPI: round 12 |

***Recommendations Specific to Meet DoD Purposes***

DoD has three main objectives from the administration of assessments to respondents as part of the broad NLSY26 effort: (a) providing updated norms; (b) informing recruiting policies; and (c) assessing score-change trends. It will be important for DoD to assess carefully the possibility of shifts in the score scale since the last norming study as DoD decides whether to rescale the ASVAB. Score-change trends cannot accurately be assessed if a significant score shift has taken place since the ASVAB norms were last set, although norms can still be updated and information relevant to recruiting policy can still be obtained. Furthermore, a misaligned score scale will lead to score misinterpretations that may be detrimental. On the other hand, although periodic rescaling should be a routine part of every testing program (see Dorans, 2002), doing so tends to upset stakeholders and to momentarily disrupt the normal flow of operations. The panel recommends that DoD evaluate any scale shifts and then carefully weigh the costs and benefits of future score-scale changes.

Because of this sensitivity to context effects, DoD is likely to want to mirror this same process for PAY2026: Administer the full operational ASVAB battery, in the same subtest order, using the same screen display, with the same limits and test lengths, etc. BLS should monitor and coordinate DoD plans on future changes to the ASVAB (i.e., the addition of Coding Speed, Cyber, Mental Counters, and Complex Reasoning tests) to ensure these are included in the NLSY2026. If marked changes to the ASVAB occur, either before the NLSY26 cohort first takes the ASVAB or before a later administration, simultaneous efforts should be made to conduct a study to assess and document the comparability of old and new ASVAB test instruments and to provide adjusted ASVAB scores that are linked across old and new versions of the ASVAB. This will allow for research assessing the stability of scores over time for NLSY26 respondents and for research examining changes in the distribution of scores (and their implications) across NLSY cohorts (Segall, 1997; Altonji et al., 2012).

There are also benefits to DoD to obtain norming data on TAPAS. Since TAPAS facet scores are influenced by which facets are included in the battery, the norming would benefit from a version of TAPAS that was used jointly by all Services. There are efforts currently underway to develop a joint-Service TAPAS battery that will contain a set of facets used by all Services. This version would be a good candidate for inclusion in the NLSY2026.

FYI scores on each of the RIASEC are translated to percentile scores based on a nationally representative sample of high schools. Students taking the FYI are encouraged to focus on the top scores (interest areas) based on these percentile scores to explore jobs with similar RIASEC emphases. For many of the same reasons ASVAB could benefit from a renorming (i.e., datedness of current norms), DoD has an interest in developing new norms for FYI based on the NLSY2026.

**1. DoD Need for a Timely Sample of 18-23 Year Old**

Consistent with NLSY79 and NLSY97, the panel recommends that a separate 18-23 year old sample be drawn for DoD purposes as part of the Household Survey for the creation of the NLSY26, rather than just sampling the younger NLSY26 cohort and waiting until it ages into the 18-23 year age range. Obtaining updated ASVAB norms on a timely basis is important for DoD as these are used for military entry. There are Congressional mandates regarding eligibility to serve (e.g., an AFQT score at or above the 10$^{th}$ percentile is a requirement for enlistment; hence the need for an accurate assessment of the score representing the 10$^{th}$ percentile). There are also targets that vary across services regarding the percent of enlistees at or above various percentile points, also driving the need for accurate norms.

While there is a current expectation that the version of ASVAB administered in the new NLSY will be comparable to the version used in 1997, should there be changes (e.g., a new score scale), we recommend that research and processes be undertaken as early as possible to establish a concordance between old and new scores.

**2. DoD Need for a Timely Sample of 10$^{th}$-12$^{th}$ Graders**

DoD's Career Exploration Program needs ASVAB and Find Your Interests (FYI) norms for 10$^{th}$-12$^{th}$ graders. Administering ASVAB and FYI to this cohort would result in a second normed sample for ASVAB for this age range for career planning purposes, to accompany the norms for 18-23 year old used for military enlistment purposes.

The panel sees at least three possible mechanisms for obtaining a 10$^{th}$-12$^{th}$ grade sample. First, and preferred by DoD, is to obtain a third separate contemporaneous sample of 10$^{th}$-12$^{th}$ graders, in addition to the NLSY main sample and the 18-23 year old sample used for operational ASVAB and TAPAS norming. The rationale for this choice is DoD's interest in making new 10$^{th}$-12$^{th}$ grade and 18-23 year old ASVAB norms at the same time.

A second option is to administer the ASVAB and FYI when respondents in the NLSY sample age into this grade range. For example, assuming the new NLSY cohort is chosen to include ages 12-16 (as was done for NLSY97), the oldest members of the cohort would be in the relevant grade range in the first round of data collection and so could take the ASVAB and FYI in the first round of data collection, while younger respondents would have to take the ASVAB and FYI in later administrations. While clearly less costly than a separate sample, two concerns about this strategy are the potential effects on sample

representativeness of attrition from the sample prior to aging into the grade range and the delay in the speed of obtaining updated 10th-12th grade ASVAB norms. This approach does not meet DoD's need for timely renorming.

A third option is to broaden the main NLSY sample from a 12-16 age range to a 12-18 age range, thus permitting ASVAB and FYI administration to the 10th-12th grade (age 16-18) portion of the main sample. In order for the sample size of 10th-12th graders to be sufficient for DoD norming, it may be that oversampling is required for this age range at initiation of the sample. This oversample could then be dropped in later NLSY rounds if cost to BLS is an issue, or surveyed at less frequent intervals (for example, once in early adulthood, once in middle age, and once at retirement age). We see this possibility as not only meeting DoD needs for timely norms, but also meeting BLS and researcher interests, as outcomes in higher education, family formation, and labor market outcomes will become available more quickly for the older portion of this sample (even if an oversample is dropped).

***Recommendations Specific to Meet BLS/Research Community Purposes***

**1. Age Range of NLSY26 at Survey Administration Initiation**

In order to consider the administration of cognitive ability, personality, and career interest assessments to the NLSY26, it is necessary to first consider the age range of the sample at initiation of survey administration. The age ranges that we have discussed include the original NLSY79 age range (14-21 on the day the sample was drawn, 14-22 when survey administration began a few months later); the original NLSY97 age range (12-16 on the day the sample was drawn, 12-17 when survey administration began a few months later); several extensions that overlap these, including 12-21 and 10-21; and a much broader age range with missing ages, such as ages 6, 9, 12, 15, 18, 21, 24. In each case, we recognize that the survey respondents will age, so that future ages will be represented in data collection (though ages younger than the youngest age will not be). A concern is that potential attrition could mitigate the value of test administration later than the first round, though we also note that early attrition in past NLSY administration has been remarkably low.

We recommend age ranges approximating the NLSY97 ages at initial administration, 12-16, though we also recommend consideration of extending the top age to 18, so that high school ages are completely represented. As discussed above, extending the age range will allow for earlier administration of the ASVAB and FYI to NLSY respondents at the DoD age of administration and allow for earlier collection of young adult outcomes in later rounds. In addition, expanding the age range has the added benefit of increasing the number of sibling pairs in the sample. Conversely, we believe that a more expansive developmental dataset with youngest ages below age 12 would be prohibitively expensive and logistically challenging (with multiple question sets that would have to be written for different ages), and the spaced ages would compromise ability to study closely-spaced cohort effects. Further, there is another way to achieve the developmental goal, illustrated by the NLSYC/YA dataset, in which children of the NLSY datasets are placed into their own dataset to support much broader developmental research (although such a sample would only be available many years in the future).

We also view the recommended age ranges above as supporting the important principle of comparability of test scores across cohorts. This principle underlines the importance of defining an "NLSY ASVAB standard" (and the same with other instruments) that can be maintained in the face of potential revision and re-norming. Compatibility with previous NLSY surveys would support using the earlier version of the instrument, whereas the revisions would be responsive to more modern testing processes and principles and would therefore be preferable on those grounds.

**2. First Administration of the ASVAB, a Personality Assessment, and FYI to the NLSY Sample**

The panel recommends administration the ASVAB and FYI to the NLSY26 when they are at an age that is appropriate for administration. In contrast, the panel does not recommend that TAPAS be administered to the NLSY26 sample. As noted in Section 2 above, TAPAS is designed specifically for use in military enlistment. The interests of non-military researchers are more likely to be aligned with frameworks such as the Big 5 personality dimensions. A Big 5 instrument (Ten Item Personality Inventory (TIPI)) was administered as part of NLSY97. We recommend continued use of TIPI with the NLSY26 sample for continuity, rather than the use of TAPAS with this cohort.

Assuming an NLSY26 cohort that is no younger than age 12, and no older than age 18, the panel recommends the following for the first administration of the ASVAB, TIPI, and FYI:

Given the considerations discussed in Section 3, the panel recommends that for the NLSY2026, individual difference measures be administered to respondents when they age into the $10^{th}$-$12^{th}$ grade. However, the panel recognizes that problems of expected or realized attrition may outweigh the benefit of waiting to administer. In addition, as noted in the previous chapter, earlier administration aids comparability with the NLSY97 and helps to avoid ceiling effects. If the BLS wishes to consider administration of these instruments to the entire sample in the first round of data collection, it should—in partnership with DoD—first conduct a reliability study as was conducted for NLSY79 for the ASVAB and the Interest Finder (McBride et al., 1999).

**3. Repeat Administration of ASVAB and Personality Measures to the NLSY Sample**

The panel sees great value in obtaining longitudinal information about various dimensions of ability and personality, and advocates various types of testing. Both domains have been linked to a broad array of high education, workplace, military, and broad life outcomes (e.g. health, marital success; Roberts et al., 2007). There is a long tradition of research on the stability of ability measures over time (e.g., Deary et al., 2000) and a recognition that a measure of general cognitive ability can be derived from ASVAB (Frey & Detterman, 2004).

There is a growing body of work on stability versus change in the personality domain. This includes looking at broad changes over the lifespan (Roberts et al., 2006), changes due to targeted interventions (Roberts et al., 2017), and changes due to specific experiences, such as experiencing military training

(Jackson et al., 2012). Re-administration of the ASVAB and personality measures to NLSY respondents will allow researchers to harness the richness of the NLSY longitudinal information in order to contribute meaningfully to this work.

Operationally, we recommend the following: First, it would be useful to re-administer ASVAB to those tested in the 10th-12th grade after age 20 (to permit time for changes to be observed) but before age 24. This will allow for research on the stability of scores across the two age categories for which DoD routinely administers the ASVAB. Second, re-administration of ASVAB periodically (e.g., age 30, age 40) would have great research value. Two existing studies provide an example of the usefulness of the re-administration of the AFQT using a small sample of Vietnam Era veterans from the VETSA sample (Lyons et al., 2017, and Eglit et al., 2022). For re-administrations up to age 30, we recommend administration of the full ASVAB, as the age range for military enlistment extends to age 32, and the ASVAB is used for re-classification among enlistees at older ages. After age 30, we recommend administering the smaller set of ASVAB subtests making up the Armed Forces Qualifying Test (AFQT) as the AFQT is a good representation of developed general cognitive ability. Third, in the personality domain we recommend re-administration of TIPI in these later test re-administrations.

## Tradeoffs that informed the ranking of recommendations

Our panel was focused on DoD-related assessments. Administering these assessments as part of the NLSY26 effort relies on assumptions of DoD support and cooperation with BLS. As a result, the panel did not think it appropriate to assess tradeoffs for the fundamental recommendations it has made. There are some minor discussions of tradeoffs embedded in our recommendations (e.g. when to administer the ASVAB for the first time to a new NLSY26 cohort), but they are minor details relative to the major recommendations we have made. The panel did not think these minor discussions merited inclusion in a separate section here.

# References

Altonji, J. G., Bharadwaj, P., & Lange, F. (2012). Changes in the characteristics of American youth: Implications for adult outcomes. *Journal of Labor Economics*, 30(4), 783-828.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). Standards for educational and psychological testing. American Educational Research Association.

Andreou, V., Peters, S., Eggermont, J., Wens, J., & Schoenmakers, B. (2021). Remote versus on-site proctored exam: comparing student results in a cross-sectional study. *BMC Medical Education*, 21(1), 1-9.

Arcidiacono, P. (2004). Ability sorting and the returns to college major. *Journal of Econometrics*, 121(1-2), 343-375.

Arcidiacono, Peter, Hotz, V. Joseph, and Kang, Songman (2012). Modeling college major choices using elicited measures of expectations and counterfactuals. *Journal of Econometrics*, 166 (1):3-16. https://doi.org/10.1016/j.jeconom.2011.06.002.

Arcidiacono, Peter, Hotz, V. Joseph, Maurel, Arnaud, and Romano, Teresa (2020). Ex Ante Returns and Occupational Choice. *Journal of Political Economy*, 128(12):4475-4522. https://doi.org/10.1086/710559.

Aughinbaugh, A., Pierret, C. R., & Rothstein, D. S. (2015). The National Longitudinal Surveys of Youth: research highlights. *Monthly Labor. Review*, *138*, 1.

Center, D. M. D. (2004). ASVAB Norms for the Career Exploration Program.

Cherry, G., O'Leary, M., Naumenko, O., Kuan, L. A., & Waters, L. (2021). Do outcomes from high stakes examinations taken in test centres and via live remote proctoring differ?. *Computers and Education Open*, 2, 100061.

Dahlke, J. A., Kostal, J. W., Sackett, P. R., & Kuncel, N. R. (2018). Changing abilities vs. changing tasks: Examining validity degradation with test scores and college performance criteria both assessed longitudinally. *Journal of Applied Psychology*, 103(9), 980.

Deary, I. J., Whalley, L. J., Lemmon, H., Crawford, J. R., & Starr, J. M. (2000). The stability of individual differences in mental ability from childhood to old age: Follow-up of the 1932 Scottish Mental Survey. *Intelligence*, 28(1), 49-55.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement*, *39*(1), 59-84.

Eglit, G. M., Elman, J. A., Panizzon, M. S., Sanderson-Cimino, M., Williams, M. E., Dale, A. M., Eyler, L. T., Fennema-Notestine, C., Gillespie, N. A., Gustavson, D. E., Hatton, S. N., Hagler, D. J., Hauger, R. L., Jak, A. J., Logue, M. W., McEvoy, L. K., McKenzie, R. E., Neale, M. C., Puckett.

O., Reynolds. C. A., Toomey, R., Tu, X. M., Whitsel, N., Xian, H., Lyons, M. J., Franz, C. E., & Kremen, W. S. (2022). Paradoxical cognitive trajectories in men from earlier to later adulthood. *Neurobiology of Aging*, *109*, 229-238.

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science*, 15(6), 373-378.

Hoff, K. A., Briley, D. A., Wee, C. J. M., & Rounds, J. (2018). Normative changes in interests from adolescence to adulthood: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 144(4):426–451. https://doi.org/10.1037/bul0000140.

Hoffman, E. P. (1987). Determinants of youth's educational and occupational goals: Sex and race differences. *Economics of Education Review*, *6*(1), 41-48.

Holland, J. L. (1997). Making vocational choices: A theory of vocational personalities and work environments. Psychological Assessment Resources.

Jackson, J. J., Thoemmes, F., Jonkmann, K., Lüdtke, O., & Trautwein, U. (2012). Military training and personality trait development: Does the military make the man, or does the man make the military?. *Psychological Science*, 23(3), 270-277.

Kim, S., & Walker, M. (2021). Assessing Mode Effects of At-Home Testing Without a Randomized Trial. ETS Research Report Series, 2021(1), 1-21.

Levine, R., & Rubinstein, Y. (2017). Smart and illicit: who becomes an entrepreneur and do they earn more?. *The Quarterly Journal of Economics*, *132*(2), 963-1018.

Levine, P. B., & Zimmerman, D. J. (1995). A comparison of the sex-type of occupational aspirations and subsequent achievement. *Work and Occupations*, *22*(1), 73-84.

Light, A., & McGee, A. (2015). Employer learning and the "importance" of skills. *Journal of Human Resources*, 50(1), 72-107.

Lyons, M. J., Panizzon, M. S., Liu, W., McKenzie, R., Bluestone, N. J., Grant, M. D., ... & Xian, H. (2017). A longitudinal twin study of general cognitive ability over four decades. *Developmental Psychology*, *53*(6), 1170.

McBride, J.R., Waters, B.K., Kaplowitz, E., Zimowski, M., Greene, H., Blair, J., Curran, L., Quenette, M (2000). Profile of American Youth 1997: Psychometric Test Administration in the NLSY97 Pretest. DMDC FR-11-10. Monterey, CA.

McBride, J.R., Waters, B.K., Stawarski, C.A., DelaRosa, M.R. (1999). Profile of American Youth 1997: Appropriateness of the Armed Services Vocational Aptitude Batter for Children Aged 12 to 14 in the National Longitudinal Survey of Youth 1997. DMDC TR-99-003. Seaside, CA.

Official website of the Armed Services Vocational Aptitude Battery. "ASVAB Fact Sheet." Accessed June, 2022. https://www.officialasvab.com/applicants/ fact-sheet/.

Prada, M. F., & Urzúa, S. (2017). One size does not fit all: Multiple dimensions of ability, college attendance, and earnings. *Journal of Labor Economics*, 35(4), 953-991.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4), 313-345.

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin*, 143(2), 117.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1), 1.

Segall, D. (2016, July 7-8). *PiCAT/ Vtest Update: Unproctored Pre-Screen with Proctored Verification Test* [Conference Presentation]. Meeting of the Defense Advisory Committee on Military Personnel Testing, Colorado Springs, CO.

Segall, D. O. (1997). "Equating the CAT-ASVAB". In W.A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association, 181-198.

Speer, J. D. (2017). Pre-market skills, occupational choice, and career progression. *Journal of Human Resources*, 52(1), 187-246.

Stoll, G., Rieger, S., Lüdtke, O., Nagengast, B., Trautwein, U., & Roberts, B. W. (2017). Vocational interests assessed at the end of high school predict life outcomes assessed 10 years later over and above IQ and Big Five personality traits. *Journal of Personality and Social Psychology*, 113(1):167–184. https://doi.org/10.1037/pspp0000117.

Weiner, J. A., & Hurtz, G. M. (2017). A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology*, 18(1), 13-20.

Wiswall, Matthew and Zafar, Basit (2021 May 5). Human Capital Investments and Expectations about Career and Family. *Journal of Political Economy*, 129(5):1361-1424. https://doi.org/10.1086/713100.

Zafar, B. (2013). College major choice and the gender gap. *Journal of Human Resources*, 48(3), 545-595.