SLIDES ONLY

NO SCRIPT PROVIDED

OPA
OFFICE OF PEOPLE ANALYTICS

# Considerations Regarding Renorming the Armed Services Vocational Aptitude Battery (ASVAB)

Rod McCloy
*Human Resources Research Organization*

Briefing presented to the DACMPT
June 13, 2024

24-P-0549

# HumRRO Colleagues/Contributors

- Steve Ferrara
- Erin McLenagan
- Cathedia Rose
- Peter Ramsberger
- Scott Oppler

# Briefing Agenda

- Background Information
- Q1: Is there evidence that renorming of the ASVAB is needed?
- Q2: What impact would renorming the ASVAB have?
- Q3: What options are available for developing new norms for the ASVAB?
- Q4: If a norming study is done, what are the options for doing so in a cost-effective and rigorous manner?
- Q5: How can the need for new ASVAB norms be assessed on an ongoing basis, and what would be the criteria for deciding renorming is necessary?
- Conclusions

# Background Information

OPA

OFFICE OF PEOPLE ANALYTICS

# Background Information

- The military Services have used ASVAB for selection/classification since 1976
- Originally scaled to data from WWII
- 1980: The test was determined to have been miscalibrated
- New norming study conducted that year
  - Data collection obtained via linking forces with the Bureau of Labor Statistics' National Longitudinal Study of Youth (NLSY)
  - The Profile of American Youth (PAY80), adjusted to maintain approximate level of expected performance of the WWII scale
- A similar norming study was conducted in 1997 (PAY97)
  - Spurred by improved AFQT scores for racial/ethnic groups
  - Collected data for both the Enlisted Testing Program and the Career Exploration Program
  - Estimated cost: $15M ($29M in 2023 dollars)

# Is It Time for Another Norming Study?

- Over time, various stakeholders have raised the question of whether a new norming study should be conducted, given that this has not been undertaken since 1997
- New norms → new composite scores and cut scores
- Potential "side effects"
  - Different score meanings associated with current benchmark scores
  - Different cut scores needed to maintain current qualification rates
  - Need to revalidate the battery

# Our Approach

- Gathered/analyzed data from multiple sources (e.g., contemporary test score trends, historical information pertinent to the questions)
- Reviewed past norming studies to identify issues and ways to circumvent them
- Developed a monitoring tool that allows for ongoing examination of changes in standardized test scores and population demographics that can be used in considering whether a new norming study is needed
- Convened a Technical Working Group (TWG) of experts in survey research, sampling methods, labor market economics, psychometrics, and educational measurement
  - Consider the pros and cons of renorming and recommend a way forward
  - Evaluate methods for monitoring the need for renorming

# Q1: Is there evidence that renorming of the ASVAB is needed?

OPA

OFFICE OF PEOPLE ANALYTICS

# The Issue

- Some stakeholders believe the ASVAB should be renormed
  - Age of data
  - Some believe it will increase applicants' eligibility for service, thereby easing recruiting difficulties
- We examined National Assessment of Educational Progress (NAEP) Long-Term Trend and National NAEP performance data to evaluate academic achievement of 18–23-year-olds since 1997 (norming)
  - Also examined 4th- and 8th-grade NAEP (future candidates)
- **BLUF: New national norms seem unnecessary at this time and will not increase eligibility**

# What We Found: NAEP Long-Term Trend Data

- Performance of 17-year-olds held steady from 1996–2012 (most recent data), with perhaps some small improvement in both reading and mathematics
- For 4th- and 8th-graders:
  - Average scores improved up until 2020 (before the upset to schooling during the COVID-19 pandemic) and then dropped off in 2022 (COVID-19 effect)

## Table 3. NAEP Long-Term Trend Mean Scale Scores (and SDs) for Years Relevant to ASVAB Norming

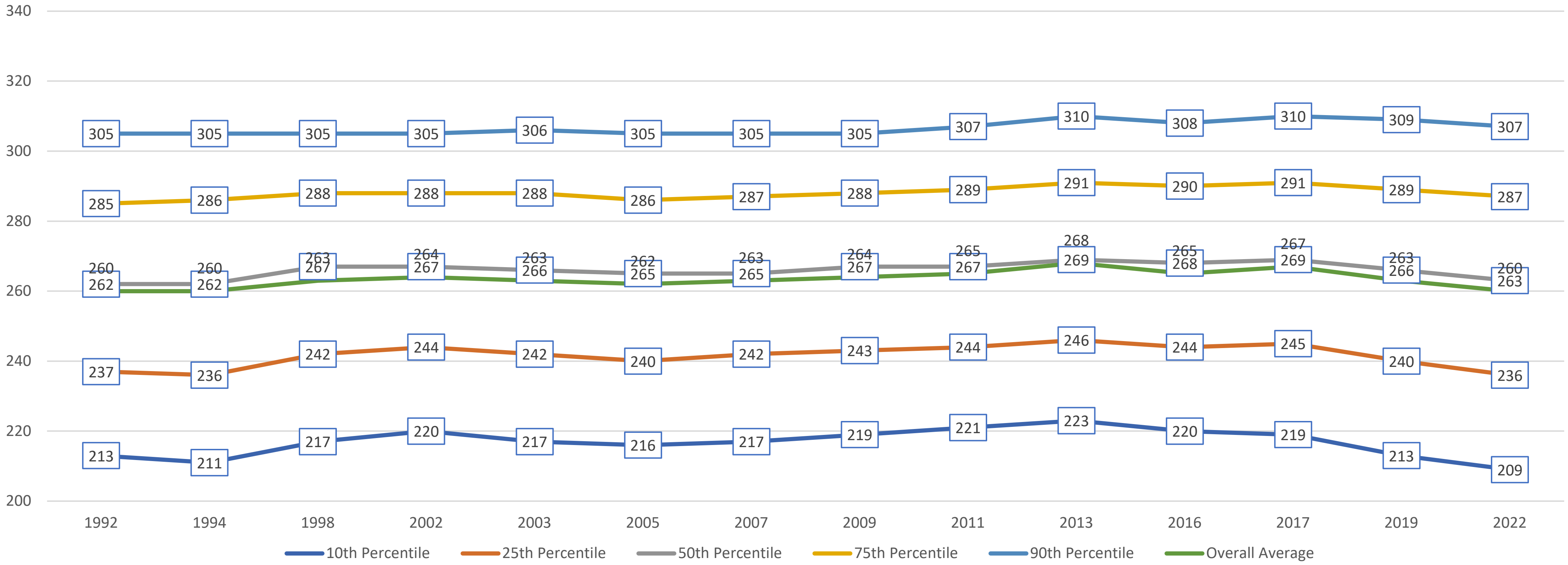
| | 1996 | 2020 | 2022 |
|---|---|---|---|
| **Reading** | | | |
| **Age 13** | 258 (39) | 260 (40) | 256 (40) |
| **Age 9** | 212 (39) | 220 (40) | 215 (43) |
| **Peak performances: 13-year-olds (263 [37]), 9-year-olds (221 [38]), both in 2012** | | | |
| **Mathematics** | | | |
| **Age 13** | 274 (32) | 280 (40) | 271 (43) |
| **Age 9** | 231 (34) | 241 (38) | 234 (41) |
| **Peak performances: 13-year-olds (285 [35]), 9-year-olds (244 [36]), both in 2012** | | | |

Values given are Mean (SD)

# What We Found: National NAEP

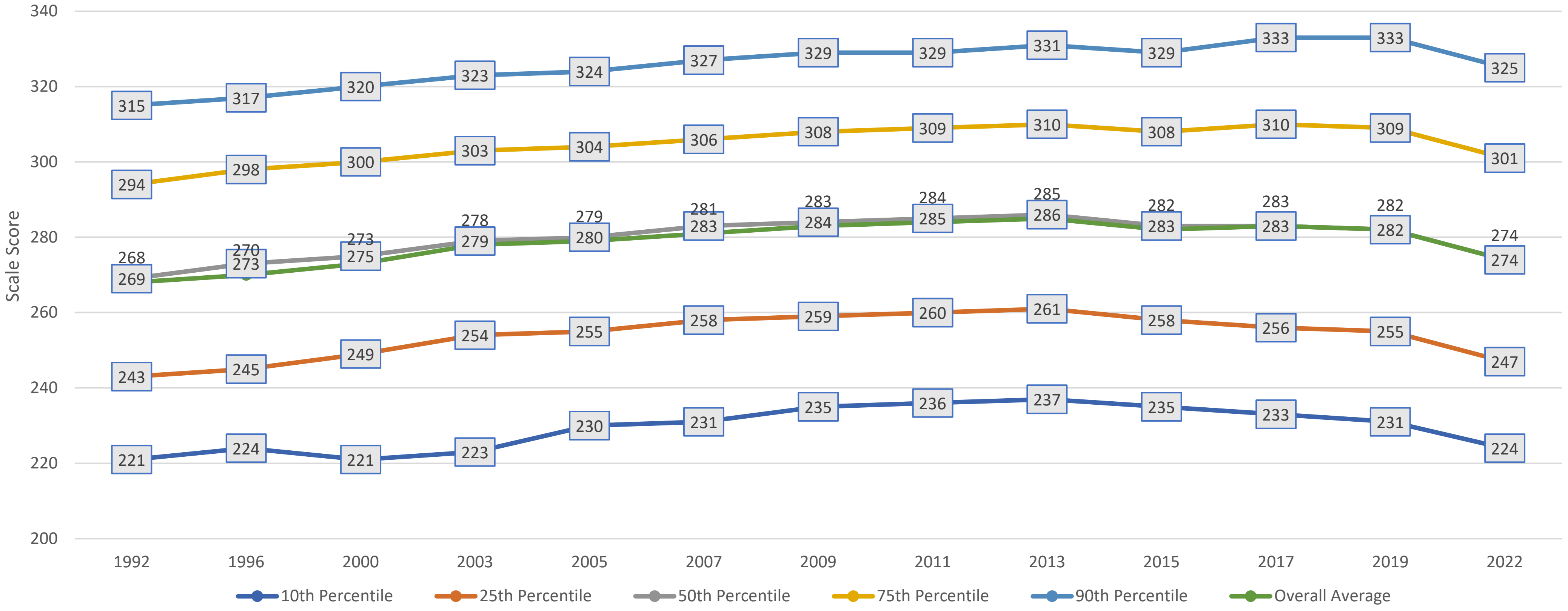
- Declines in performance for grade 8 reading have been occurring at the lower levels of achievement—primarily the 10th and 25th percentiles—and at the 50th percentile
  - This trend at the lower levels of achievement appears in the other grades in reading (since 2017) and in mathematics (since 2013) (see Figures 1 and 2, slides 13 and 14)
- Student achievement increased in grade 4 reading between 1998 and 2019 (unlike the Long-Term Trend results) and declined in grades 8 and 12 during the same period (see Table 4, slide 15)
  - Thus, reading performance was declining prior to the pandemic
- Grade 12 mathematics performance remained the same between 2005 and 2019 and increased at the lower grades by four points (grade 8) and 12 points (grade 4) between 1996 and 2022 (see Table 4, slide 15)

# Figure 1. National NAEP Grade 8 Reading Average Scores at Selected Percentiles, 1992-2022



From https://www.nationsreportcard.gov/reading/nation/scores/?grade=8

# Figure 2. National NAEP Grade 8 Math Average Scores at Selected Percentiles, 1992-2022



From https://www.nationsreportcard.gov/mathematics/nation/scores/?grade=8

Reading performance was declining prior to the pandemic for older students but improving for 4th-graders.

## Table 4. National NAEP Mean Scale Scores (and Standard Deviations) for Years Relevant to ASVAB Norming

| | Mean | SD | Mean | SD | Mean | SD |
|---|---|---|---|---|---|---|
| **Reading Grade** | **1998** | | **2019** | | **2022** | |
| **12** | 290 | 38 | 285 | 42 | -- | -- |
| **8** | 264 | 35 | 263 | 38 | 260 | 38 |
| **4** | 215 | 39 | 220 | 39 | 217 | 40 |
| **Peak performances: 12th-graders [290 (38)] in 1998; 8th-graders [268 (34)] in 2013; and 4th-graders [223 (37)] in 2015** | | | | | | |
| **Mathematics Grade** | **2005** | | **2019** | | **2022** | |
| **12** | 150 | 34 | 150 | 36 | -- | -- |
| | 1996 | | 2019 | | 2022 | |
| **8** | 270 | 37 | 282 | 40 | 274 | 39 |
| **4** | 224 | 31 | 241 | 32 | 236 | 33 |
| **Peak performances: 12th-graders [153 (34, 33)] in 2009 and 2013; 8th-graders [285 (37)] in 2013; and 4th-graders [240 (30)] in 2013** | | | | | | |

*Note*. From https://www.nationsreportcard.gov/ndecore/xplore/NDE

Mathematics performance remained constant for seniors. Younger students showed gains prior to the pandemic but losses thereafter.

# Lower-Scoring Students in the National NAEP Data

| NAEP Test | NAEP Proficiency Level | % of Students in Category |
|---|---|---|
| Reading | Below Basic | 30 |
| | Basic | 33 |
| Mathematics | Below Basic | 40 |
| | Basic | 35 |

Rough correspondence to those in AFQT Categories IV and V (30% in the norming population). Thus, the military might expect candidates scoring in these categories (percentiles 1–30) not to be able to perform at a Basic level (i.e., not to display partial mastery of reading and mathematics knowledge and skills).

# Comments from the TWG

- **ASVAB renorming is NOT called for at this time**
  - NAEP results do not support the conclusion that academic achievement has improved since implementation of the PAY97 norms in 2004
  - Little change in achievement despite large demographic changes
  - Post-pandemic drop in student scores
    - In addition to NAEP, Kuhfeld et al. (2022) examined MAP growth assessment data for 5.4 million students and found:
      - Consistent drops in scores in reading and math
      - Increases in the gap between students by SES
    - Hough and Chavez (2022) found the same for CA students on the Smarter Balanced Assessments (greater declines in lower grades)
    - Achievement score stabilization should be observed before renorming is considered
      - Affecting all students K–12, could take a decade or more to rectify itself

# Q2: What impact would renorming the ASVAB have?

OPA
OFFICE OF PEOPLE ANALYTICS

# Three Scenarios

| Scenario | Recruiting | Expected Training and Job Performance |
|---|---|---|
| 1: Lower Performance than PAY97 Norm Group | Improved | Worse |
| 2: Higher Performance than PAY97 Norm Group | Worse | Improved |
| 3: Same Performance | Unchanged | Unchanged |

# Scenario 1

- The current AFQT score from the sum of standardized subtest scores (SSSS) is 183 for the bottom of AFQT Category III-B (percentile = 31)
- If performance drops in the new norm group, the score associated with the 31st percentile will likely drop, too—say, to 178
    - Those who score from 178 to 182
        - Do not qualify as III-B under the current norms
        - Would qualify as III-B under the new norms
    - Qualification status would change (more qualify) despite their standing on AFQT not changing

NOTE: The three scenarios assume we do NOT fix the AFQT score scale but rather retain the percentile cuts.

# Scenario 2

- The current AFQT score from the sum of standardized subtest scores (SSSS) is 183 for the bottom of AFQT Category III-B (percentile = 31)
- If performance increases in the new norm group, the score associated with the 31st percentile will likely increase, too—say to 188
    - Those who score from 183 to 187
        - Do qualify as III-B under the current norms
        - Would not qualify as III-B under the new norms
    - Qualification status would change (fewer qualify) despite their standing on AFQT not changing

NOTE: The three scenarios assume we do NOT fix the AFQT score scale but rather retain the percentile cuts.

# Scenario 3

- If there is little change in AFQT performance in the new norming sample, then the status quo obtains
  - Qualification rates and expected training/job performance would be unchanged

# Comments from the TWG

- It is ***crucial*** to maintain the existing score scale
  - This would entail altering the percentile boundaries of the AFQT categories
    - If in Scenario 1 the AFQT score of 183 would indicate the 36th percentile rather than the 31st, then the AFQT score categories could be updated to preserve their meaning in the PAY97 norming sample based on their associated AFQT score calculated as the sum of standardized subtest scores.
    - Thus, Category IIIB might be changed to have a lower bound of 36 in the new norming sample. This would ensure that the same level of AFQT performance is reflected at the bottom of that category that is currently reflected in the PAY97 sample (i.e., the 31st percentile of the current norms and the 36th percentile of the new norms would both be associated with the same AFQT score of 183).
    - Similarly, the AFQT score associated with the upper bound of Category V could be increased to ensure that all those captured in that category in the PAY97 sample also were declared ineligible in the updated norms.

# Comments from the TWG (cont.)

- The opposite implications would accrue for Scenario 2
  - The upward shift in the norming sample would require decreases in the cut scores for the various AFQT categories to maintain the expected levels of performance associated with those categories
    - For example, the bottom of the III-B category might drop to the 28th percentile from the 31st as in the PAY97 sample to retain the meaning of an AFQT SSSS score of 183
- Such changes in performance standards would ensure that the same candidates qualifying on the current scale would also qualify on the new score scale
- It would reinforce the status quo, however; thus, providing no benefits in terms of recruitment prospects or expected performance

# Q3: What options are available for developing new norms for the ASVAB?

OPA
OFFICE OF PEOPLE ANALYTICS

# Five Options

- The TWG considered five options regarding renorming the ASVAB
  - Renorming
  - Applicant-Based Norming
  - Weighted Applicant-Based Norming
  - Reweighting PAY97 Data
  - Maintaining the Current Norms

# Option 1: Renorming

| Arguments For | Arguments Against |
|---|---|
| The norms would . . .<br>– look less dated to stakeholders (i.e., more current)<br>– demonstrate attention to stakeholder concerns<br>– provide an updated look at the potential pool of military eligible youth | – Changes in achievement data do not suggest renorming is needed<br>– Cost<br>– Could result in few differences between the updated and 1997 norms<br>– May need to be repeated in the near future depending on whether pandemic effects are near- or long-term<br>– Could change score scale interpretations and/or qualification rates<br>– Could make it more difficult to track recruit quality over time (i.e., if score scale interpretations change) |

# Option 2: Applicant-Based Norming

| Arguments For | Arguments Against |
|---|---|
| – No need for new data collection<br>– Supports interpretation of scores in comparison to the applicant pool<br>– Provides greater alignment between the samples upon which the norms are based and the pool of actual applicants | – Might not satisfy recruiters, especially if the number of qualified candidates does not increase<br>– Represents a change in reference point from previous norming efforts<br>– Misses an opportunity to get an updated look at the potential pool of military eligible youth<br>– Does not account for differences between the applicant population and the overall youth population, including the relatively small number of females in the former<br>– The applicant population may fluctuate from year to year in response to economic and societal factors<br>– Could make it more difficult to track recruit quality over time (i.e., if score scale interpretations change) |

Under this option, the DoD would simply report performance of current ASVAB examinees in norms based on current or recent examinees rather than that from a nationally representative sample. The ACT, GRE, LSAT, and MCAT all report candidate performance using examinee-based norms. The reporting norms for these programs are based on examinees and do not represent the national population.

# Option 3: Weighted Applicant-Based Norms

| Arguments For | Arguments Against |
|---|---|
| – No need for new data collection<br>– Provides some indication of ability level of the overall youth population | – Weights for groups underrepresented in the application population would be substantial<br>– Misses an opportunity to get an updated look at potential pool of military eligible youth<br>– Does not fully account for differences between the applicant population and the overall youth population<br>– Could make it more difficult to track recruit quality over time (i.e., if score scale interpretations change) |

This approach is similar to the preceding one with the only difference that the applicant data would be reweighted to reflect the national population of youth on key demographic variables (e.g., race/ethnicity, gender). College Board does both Options 2 and 3 for the SAT. Choosing Option 3 for the ASVAB would result in estimates of the performance of a nationally representative sample of examinees but without the expense of a nationwide data collection.

# Option 4: Reweighting PAY97 Data

| Arguments For | Arguments Against |
|---|---|
| – No need for new data collection | – Might not satisfy those who feel renorming is needed |
| – Provides indication of whether population demographic changes would have meaningful impact on AFQT estimates | – Misses an opportunity to get an updated look at the potential pool of military eligible youth |
| | – Assumes no significant change in ability levels of population subgroups |

This approach would involve reweighting the data from the PAY97 study using current population estimates from the Census Bureau on key demographic characteristics. Assuming there have been no significant changes in the ability levels of youth within the various demographic groups, this would provide evidence regarding whether demographic changes in the norming group would have a meaningful effect on AFQT estimates. The population would be stratified by age, race-ethnicity, gender, and educational attainment. Using the updated demographic information, AFQT estimates would be generated that could be compared to those from 1997. If significant differences in those estimates were found, it would suggest the need for a new norming study.

# Option 5: Maintaining the Current Norms

| Arguments For | Arguments Against |
|---|---|
| – No need for new data collection<br>– Avoids disruption<br>– Stakeholders are familiar with the current scores and their interpretation with regard to applicants’ knowledge and ability | – Would not satisfy those who feel renorming is needed given the long interval since the last effort<br>– Might lead stakeholders to believe their voices are not being heard by the Department<br>– Misses an opportunity to get an updated look at the potential pool of military eligible youth<br>– COVID-19 pandemic effects |

Over time, military manpower and training personnel have become familiar with the meaning of ASVAB scores and ability groupings and their implications. Further, numerous validity studies have been carried out that demonstrate the relations between ASVAB scores and subsequent performance in training and on the job. Although there is some value in obtaining an updated picture of the overall ability levels of American youth, the expense of doing so is difficult to justify at this time. In addition, there is evidence that the pandemic had an impact on the educational outcomes of students, and the TWG recommended waiting so that norming is not conducted on a cohort influenced by these effects.

# Q4: If a norming study is done, what are the options for doing so in a cost-effective and rigorous manner?

# Two Primary Options

- National Longitudinal Survey of Youth (NLSY) in 2026
    - DTAC has begun discussions with BLS
        - Some sense that this might not be feasible due to different timelines
- Joint Advertising, Market Research, & Studies (JAMRS)
    - JAMRS conducts several surveys to
        - Obtain data on youth propensity to serve and attitudes towards the military
        - Assess the impact of DoD and Service advertising
    - Database spans ~95% of the youth market
        - Info on gender, race, ethnicity, and geographic location (but restricted to juniors in HS and above)

# Q5: How can the need for new ASVAB norms be assessed on an ongoing basis, and what would be the criteria for deciding renorming is necessary?

# Monitoring Tool

- Excel spreadsheet
- Allows stakeholders to examine trends in test scores and other relevant data (e.g., population demographics) for consideration in deciding when/if renorming might be necessary
- Effect sizes are reported (Cohen’s *d*) as a means to determine the magnitude of any shifts in mean test performance

# External Data Sources

**NAEP Data**

- Long-Term Trend Data
  - Reading and Mathematics
  - Mean and SD across years
  - Timespan: 1971–2012 (last year of administration to 17-year-olds)
- National Data
  - Data for 1988 (closest to PAY97), 2019 (pre-COVID baseline), and 2022 (most recent)
- Lack of changes over time do not suggest renorming at this time

**SAT Data**

- Reading and Mathematics
- Have 3 years of data
  - Means and standard deviations
  - All examinees, race/ethnic groups, and gender
- Small upward changes in math scores for Blacks and Asians in recent years
- All groups show moderate reading improvements since 2008
  - Caution: change in format in 2017 (reading and writing portions combined)

**ACT Data**

- Reading and Mathematics
  - By race/ethnicity and gender
- Years 1997–2021
- Nearly all effect sizes <0.20

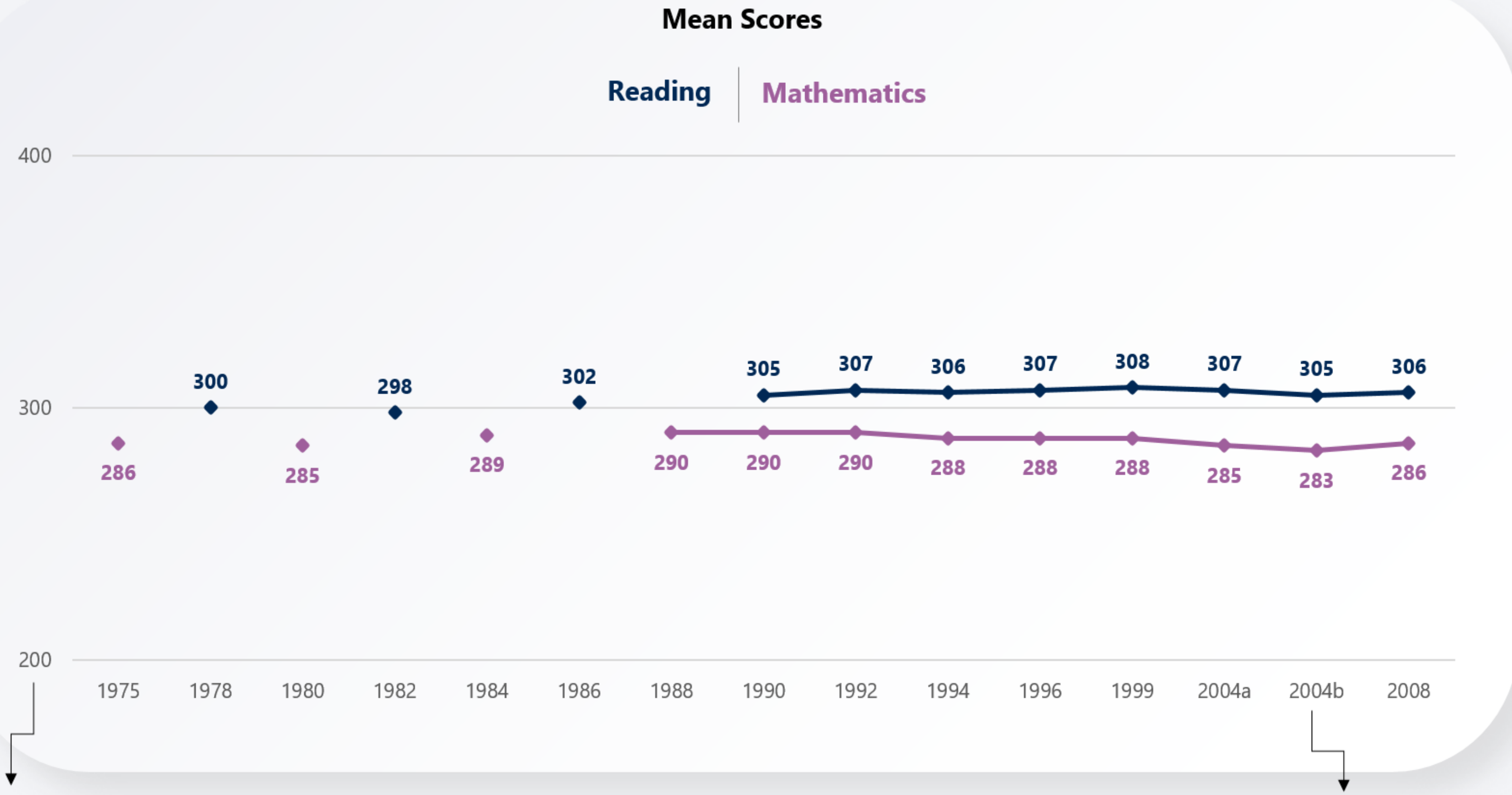
**Population Demographics**

- Race/ethnicity and gender
  - 18–23-year-olds
  - Years: 1997–2022

# NAEP Data

- Long-Term Trend Data
  - Reading and Mathematics
  - Mean and SD across years
  - Time span: 1971–2012 (last year of administration to 17-year-olds)
  - Comparisons do not suggest significant changes in recent years
- National Data
  - Data for 1998 (closest to PAY97), 2019 (pre-COVID baseline), and 2022 (most recent)
  - Lack of changes over time do not suggest renorming at this time

# ASVAB Renorming Indicator: NAEP Long-Term Trend by Content Area



## Mean Scores

full scale ranges from 0 - 500

new assessment format started

## Effect Sizes

Comparison to

| 1971 | 1996 | Recent Year |
|---|---|---|

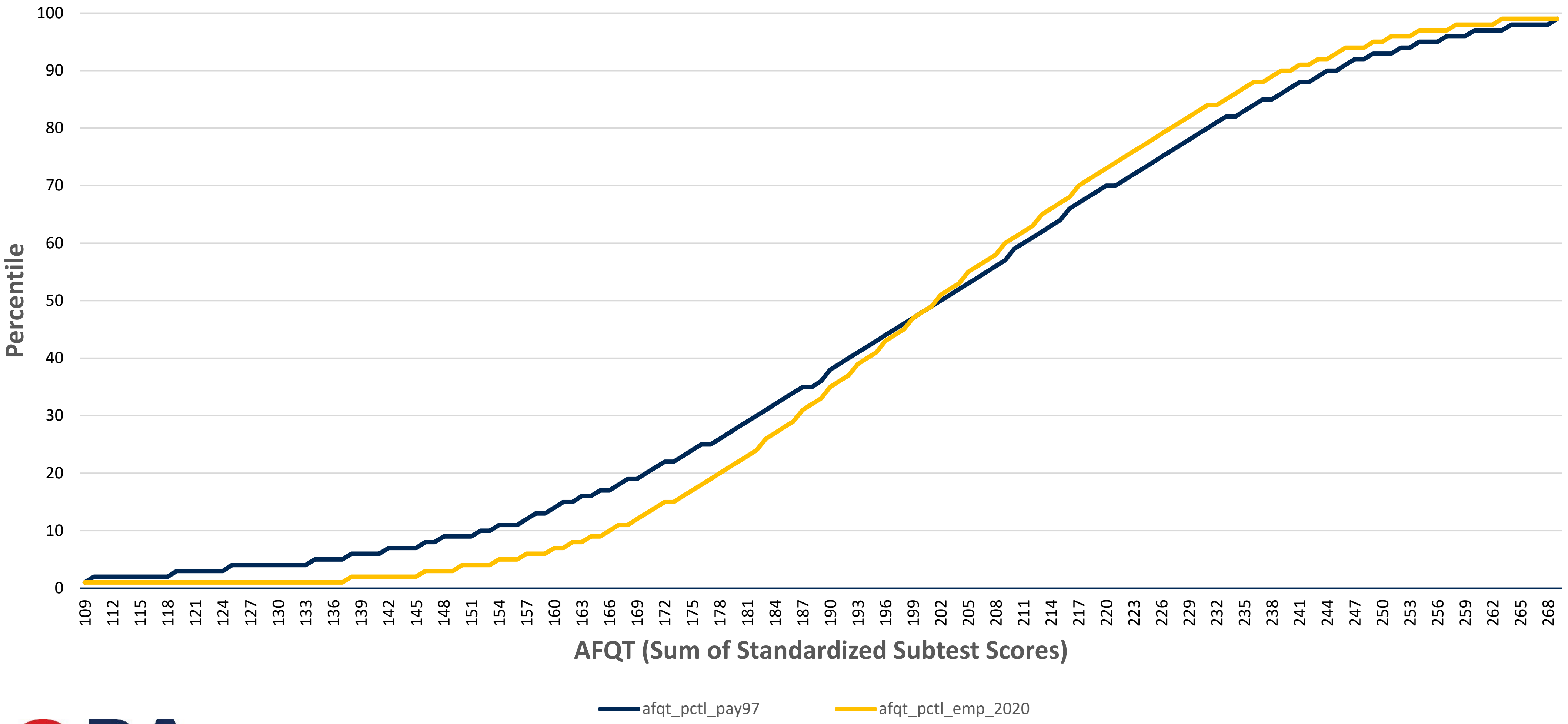
| Year | Reading | Mathematics |
|---|---|---|
| 1975 | 0.02 | |
| 1980 | | |
| 1982 | | -0.06 |
| 1984 | 0.10 | |
| 1986 | | 0.06 |
| 1988 | 0.14 | |
| 1990 | 0.12 | 0.16 |
| 1992 | 0.12 | 0.23 |
| 1994 | 0.07 | 0.20 |
| 1996 | 0.07 | 0.23 |
| 1999 | 0.07 | 0.26 |
| 2004 | | 0.23 |
| 2008 | 0.02 | 0.20 |
| 2012 | 0.05 | 0.19 |

Trivial Small Mod Large

# ASVAB Score Distributions

- Previous data provide statistical moments (means, standard deviations)
- Some TWG members suggested examining distributions instead
- Two examples—comparisons of:
  - PAY97 to 2020 ASVAB examinees
  - June2019–March2020 to April2020–March2022

# Distributions of the Sum of Standardized Subtest Scores for AFQT and Associated Percentile Scores in the PAY97 and a 2020 Applicant Sample

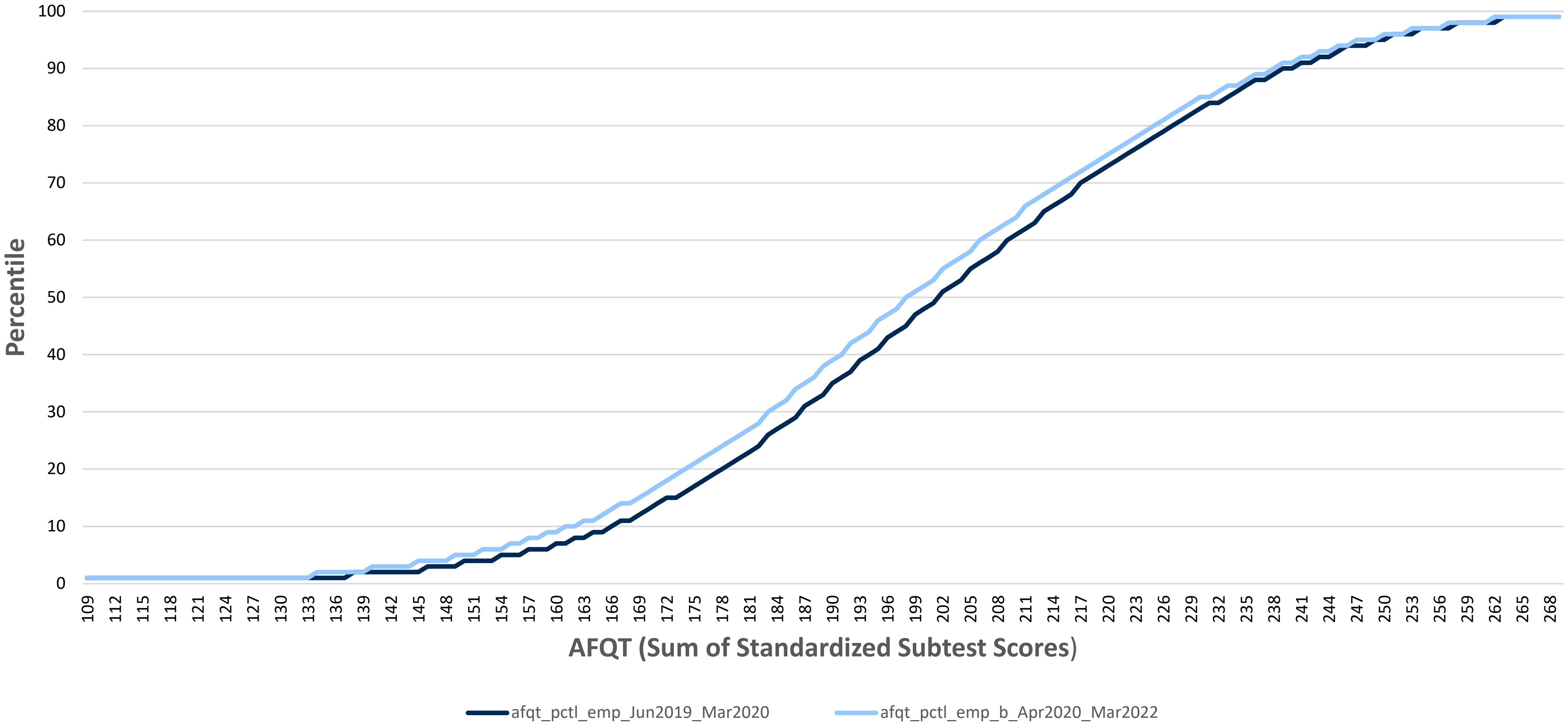


OFFICE OF PEOPLE ANALYTICS

# PAY97/AFQT 2020 Comparison

- 2020 examinee sample shows truncation at both ends of the distribution
  - Regression to the mean effect
- Interpretation is somewhat uncertain
  - How much of this discrepancy/regression to the mean in the 2020 examinee sample is due to shifts in the youth population?
  - How much is due to the slight disconnect between the PAY97 sample (which represents the national youth population but not military applicants) and the recent applicant sample (the latter being a more self-selected group)?

# Distributions of the Sum of Standardized Subtest Scores for AFQT and Associated Percentile Scores in Two ASVAB Applicant Samples



OFFICE OF PEOPLE ANALYTICS

# June2019–March2020 / April2020–March2022 Comparison

- A more direct determination of shifts in the ASVAB score distribution over time
- This approach emanates from a similar point of view as the use of applicant data to update test norms
  - Used by several high-stakes testing programs, including ACT, MCAT, and LSAT
- Shows a general decline in ability among the recent ASVAB examinees
  - The curve for examinees from **April 2020 – March 2022** is elevated slightly above the curve for examinees from June **2019 – March 2020**
  - This elevation in the more recent sample means that a given sum of standardized subtest scores for the AFQT (say, 200) represents a higher percentile score (52nd percentile) in the more current sample than in the 2019–2020 sample (48th percentile).

# June2019–March2020 / April2020–March2022 Comparison (cont.)

- This slight difference in distributions suggests renorming based on a similar ability sample might make it somewhat easier to qualify at current cut scores
  - But, this could result in a lower-performing force (similar to Scenario 1)
- Unclear how long this decline could be evidenced if induced by temporarily reduced learning during the COVID-19 pandemic

# Criteria for Deciding Whether Renorming Is Necessary

- No Single Marker
  - No single statistic—shift in K–12 student achievement, change in the demographic makeup of America’s youth, or automated algorithm—can signal to the DoD when might be the time to update the ASVAB norms
- Stats Can Guide Monitoring Activity
  - Combinations of statistics (e.g., effect sizes for achievement test scores), practical impacts (e.g., meaningful changes in distributions of AFQT scores), and meaningful demographic shifts (e.g., increases in numbers of English learners taking ASVAB) can guide the DoD in making decisions about ASVAB norms
  - Applying Cohen’s framework, we suggest users can disregard (but continue to monitor) small differences in means and proportions (e.g., < 0.5), and consider medium differences (e.g., > 0.5) as signals that updating the ASVAB norms may be necessary
- Ultimately a Policy Decision Based on Judgment
  - Human judgment plays a significant role in deciding whether a change in achievement or demographics warrants updating the norms

# A Final Thought

- The TWG members declared that the PAY97 norms have essentially become a criterion-referenced scale backed by a large amount of validity data
- Thus, even if the youth population has shifted somewhat in terms of ability, their standing is best understood relative to the current scale that has been in place for more than 25 years
    - DoD personnel managers know what to expect from recruits who score in the IIIB range vs. II
    - Renorming would reshuffle this interpretation of scores and expected performances from those attaining them
    - This disruption is arguably not worth the cost of renorming—particularly now in light of
        - Very minor shifts in scores in the youth samples of interest
        - The pending disruption from COVID-19 on youth ability profiles

# Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

# Questions for the DAC

- Are there additional options we should consider for developing new norms for the ASVAB (per Q3)?
- Can you suggest any additional criteria we should examine/monitor to serve as harbingers of the need to renorm the ASVAB (per Q5)?
- Do you have concerns with dispensing with CEP-specific norms (going only with ETP norms)? Do you anticipate a significant impact of not having grade-specific norms for the CEP?
    - JAMRS database excludes 16-year-olds
    - DTAC plans to drop gender-specific norms for the CEP

# Thank you!

For more information please contact:

**Rod McCloy**
**rmccloy@humrro.org**
**502.966.7012**

OPA
OFFICE OF PEOPLE ANALYTICS