



# **[Adverse] Impact of the ASVAB and Special Tests: Findings for Fiscal Year 2021 Applicants**

Gregory Manley

Ping Yin

Mary Pommerich

*Defense Testing & Assessment Center (DTAC)*

DACMPT Meeting  
December 16, 2022

# POTENTIAL FOR ADVERSE IMPACT

- Adverse impact (AI) is the unintended discrimination of a protected class that is the result of a selection procedure (Uniform Guidelines, 1978).
- AI is not a property of a test, per se. However, AI may occur when a test's scores are used as the bases for selection.
- A selection test may contribute potential for AI when it shows sizable mean test score differences between a majority group and a protected class (minority).
- Effect sizes of the standardized mean difference gives us an index to examine a test's potential for AI.

# HOW IS ADVERSE IMPACT ASSESSED?

- **The four-fifths rule is often used to determine the occurrence of adverse impact:**

“A selection rate for any race, sex, or ethnic group, which is less than four-fifths (80%) of the rate for the group with the highest rate, will generally be regarded by the Federal enforcement agencies as evidence of adverse impact.”

[Section 60-3, Uniform Guidelines on Employee Selection Procedures (1978); 43 FR 38295 (August 25, 1978).]

- **The ratio comparing the selection rates is called the *impact ratio*:**

$$IR = \frac{SR_{Foc}}{SR_{Ref}}, \quad \text{where } SR \text{ is the selection rate}$$

- **Ideally  $IR = 1$ , but 4/5ths leaves wiggle room**

# HOW IS ADVERSE IMPACT ASSESSED?

- **The four-fifths rule (80%) and accompanying statistics are applied to the ASVAB qualifying test (AFQT) by comparing qualification rates across the focal and reference groups of interest with regard to:**
  - Examinees who qualify for entry into the military (i.e., those scoring in AFQT category IIIB or higher,  $AFQT \geq 31$ ).
  - Examinees who qualify for enlistment incentives (i.e., those scoring in AFQT category IIIA or higher,  $AFQT \geq 50$ ).
  - Adverse impact is assessed using initial test scores only (i.e., scores from retests or confirmation tests are excluded from the analyses).
  - Significance testing is not necessarily useful analyses with very large numbers of applicants (i.e.,  $>2000$ ).

# POTENTIAL FOR ADVERSE IMPACT

- Effect sizes (ES) (i.e., standardized mean differences, AKA Cohen's  $d$ ) provide a method of evaluating potential for adverse impact across individual ASVAB and Special Tests, where no direct selection occurs.
- Effect sizes are computed for all group comparisons as:

$$ES = \frac{\mu_R - \mu_F}{\sigma_p}$$

where:

$\mu_R$  is the mean score in the Reference (Majority) group.

$\mu_F$  is the mean score in the Focal (Minority) group.

$\sigma_p$  is the pooled standard deviation across the two groups.

Note. Positive values are the direction of minority impact.

# CONFIDENCE INTERVALS ABOUT EFFECT SIZES

- A 95% confidence interval ( $\delta_L, \delta_U$ ) for the effect size (ES) is computed as (Hedges & Olkin, 1985):

$$\delta_L = ES - 1.96\hat{\sigma}(ES) \quad \delta_U = ES + 1.96\hat{\sigma}(ES)$$

where:

$$\hat{\sigma}(ES) = \sqrt{\frac{n_R + n_F}{n_R n_F} + \frac{ES^2}{2(n_R + n_F)}}$$

- Effect sizes can be plotted and classified with respect to Cohen's (1988) standards of evaluation.
  - **Small** effect sizes start at 0.20.
  - **Moderate** effect sizes start at 0.50.
  - **Large** effect sizes start at 0.80.

# WHO IS ASSESSED FOR ADVERSE

## IMPACT?

- The ASVAB testing program evaluates (adverse) impact for the following pairs of groups:

---

Pair	Reference Group	Focal Group
1	Males	Females
2	Non-Hispanic Whites	Hispanic Whites
3	Non-Hispanic Whites	Non-Hispanic Blacks
4	Non-Hispanic Whites	Non-Hispanic Asians

---

- The focal group is potentially disadvantaged relative to the reference group.
- Pairs 1–3 are the same groups that are used in evaluating DIF. Pair 4 is included since FY2017 because Non-Hispanic Asians now represent >2% of the applicant population.

# WHEN IS ADVERSE IMPACT MEASURED?

- Ideally, AI is assessed on a regular basis.
- DTAC's longitudinal analysis program examines AI for every odd-numbered FY since FY2005
  - FY2005 – 2021 odd-numbered years (excl. FY2007)
- In the current study, AI is measured for applicants testing in FY2021\*
  - (Oct 1, 2020 – Sept 30, 2021)

\*See “caution” next slide

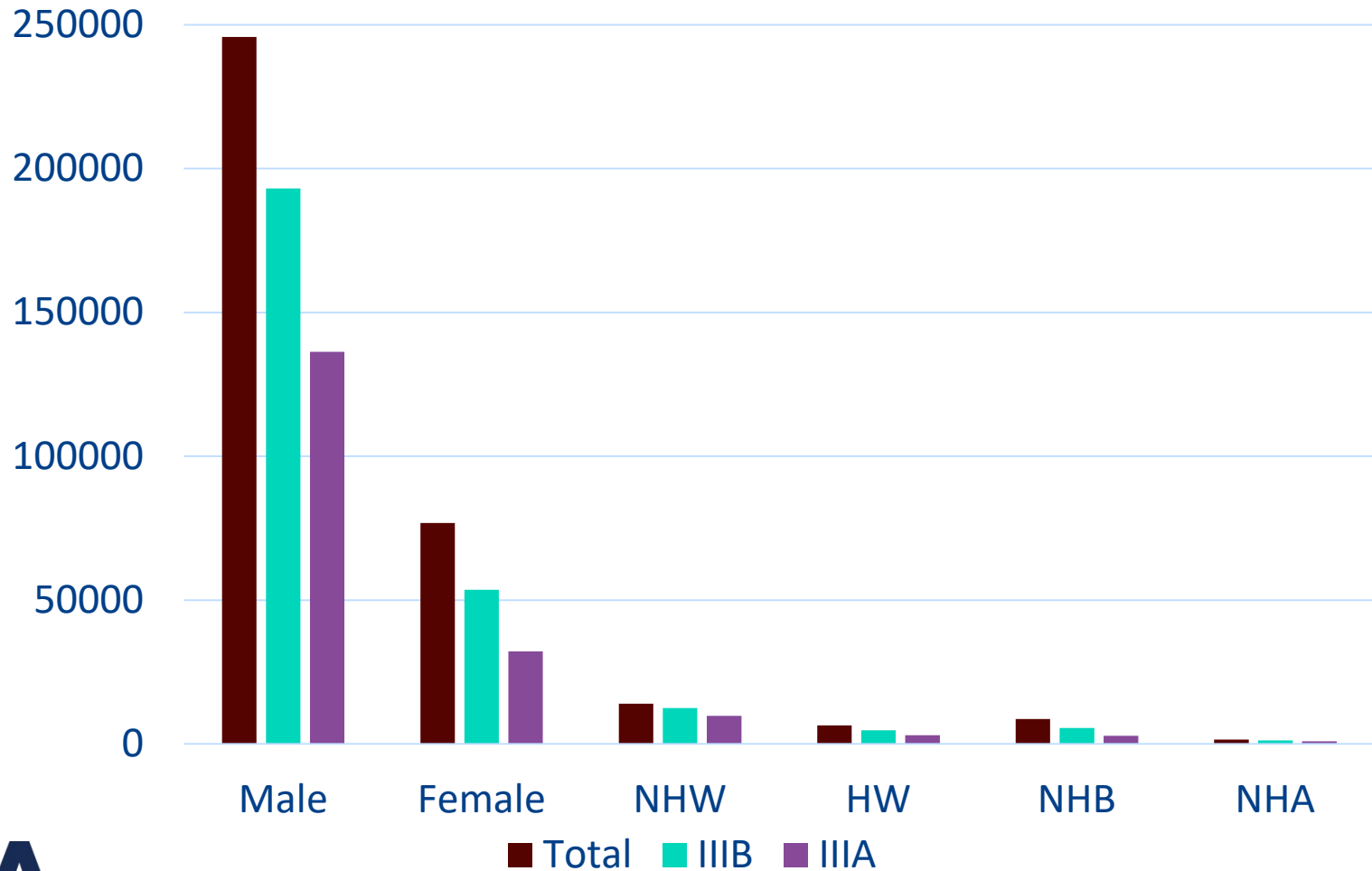


# CAUTION: DIFFERENCES BETWEEN FY2021 DATA AND PRIOR FISCAL YEARS



- FY2021 includes parts of COVID-19 “shutdown” year and the year following the shutdown. Issues with missing demographic data: Sample sizes for Race/Ethnicity are very small and may not be representative of the overall population of actual applicants. However, most effect sizes are still similar to previous years.
- Nonetheless, these analyses should provide insight to the effort for removing aptitude barriers that adversely impact diversity.

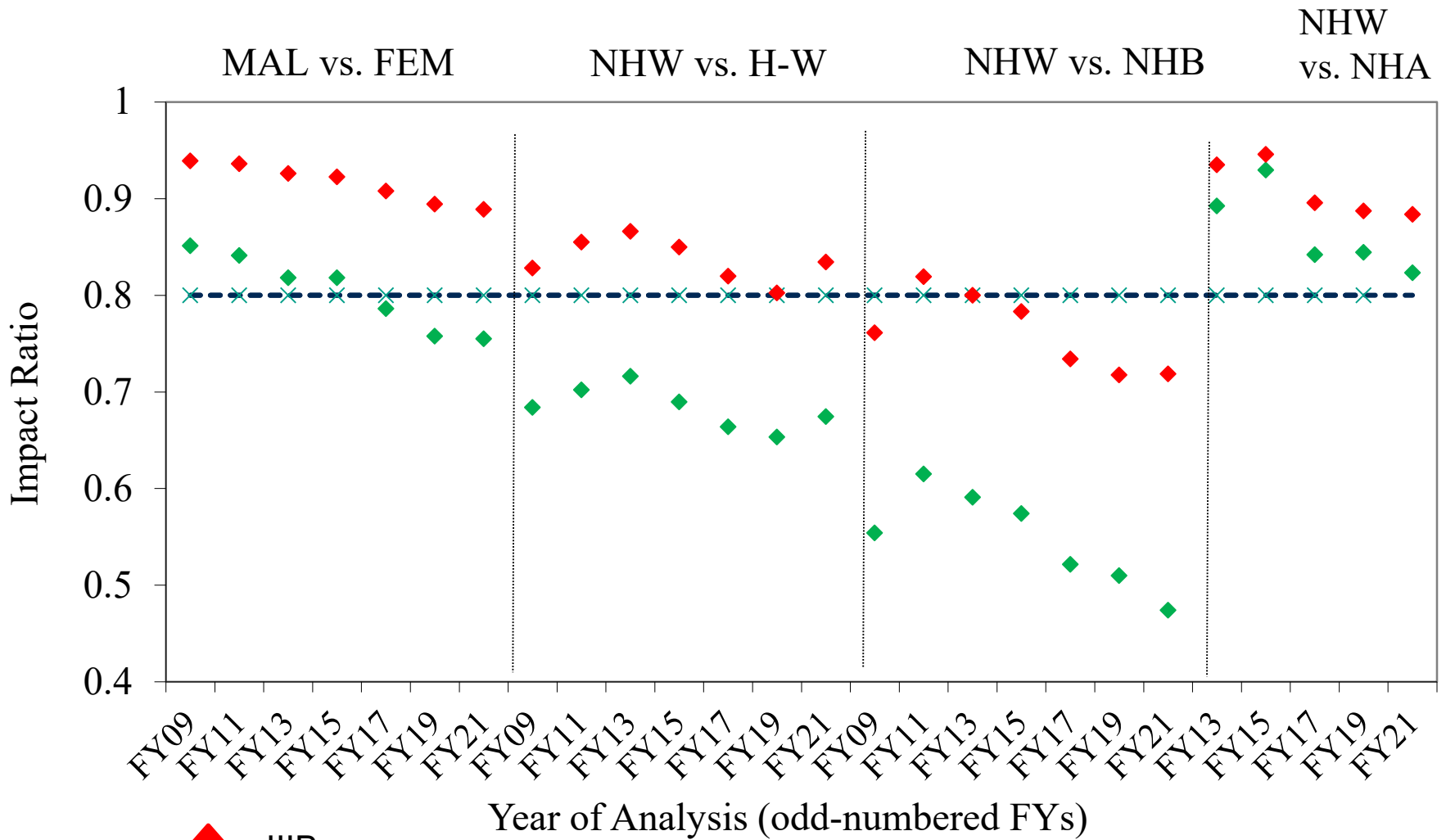
# Adverse Impact Analysis Sample Sizes for FY2021



## Impact Ratio (and 95% Confidence Interval) for AFQT Cutscores FY2021 IIIB+ & IIIA+ (All education levels)

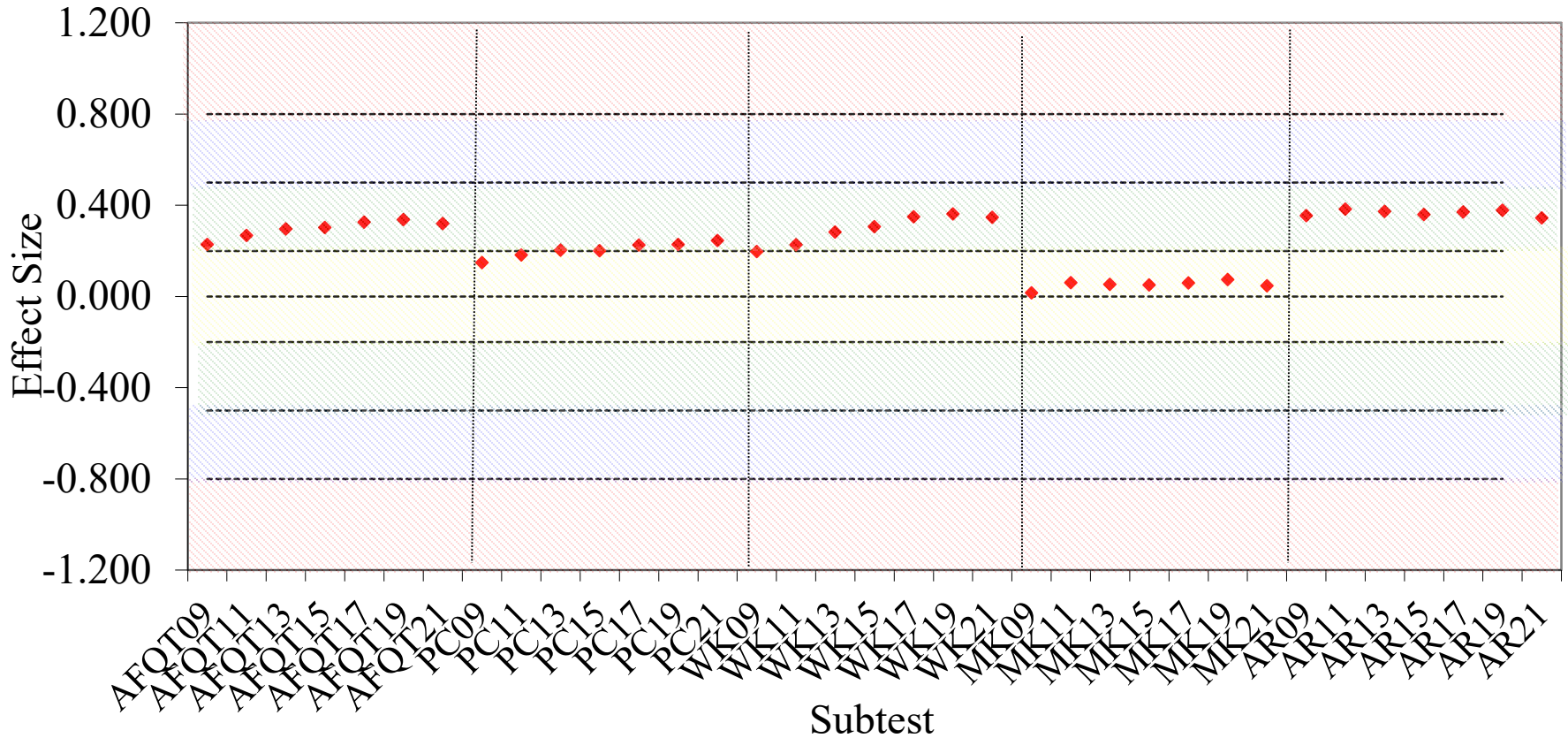


# Comparison of Impact Ratios for Odd-Numbered FYs 09-21



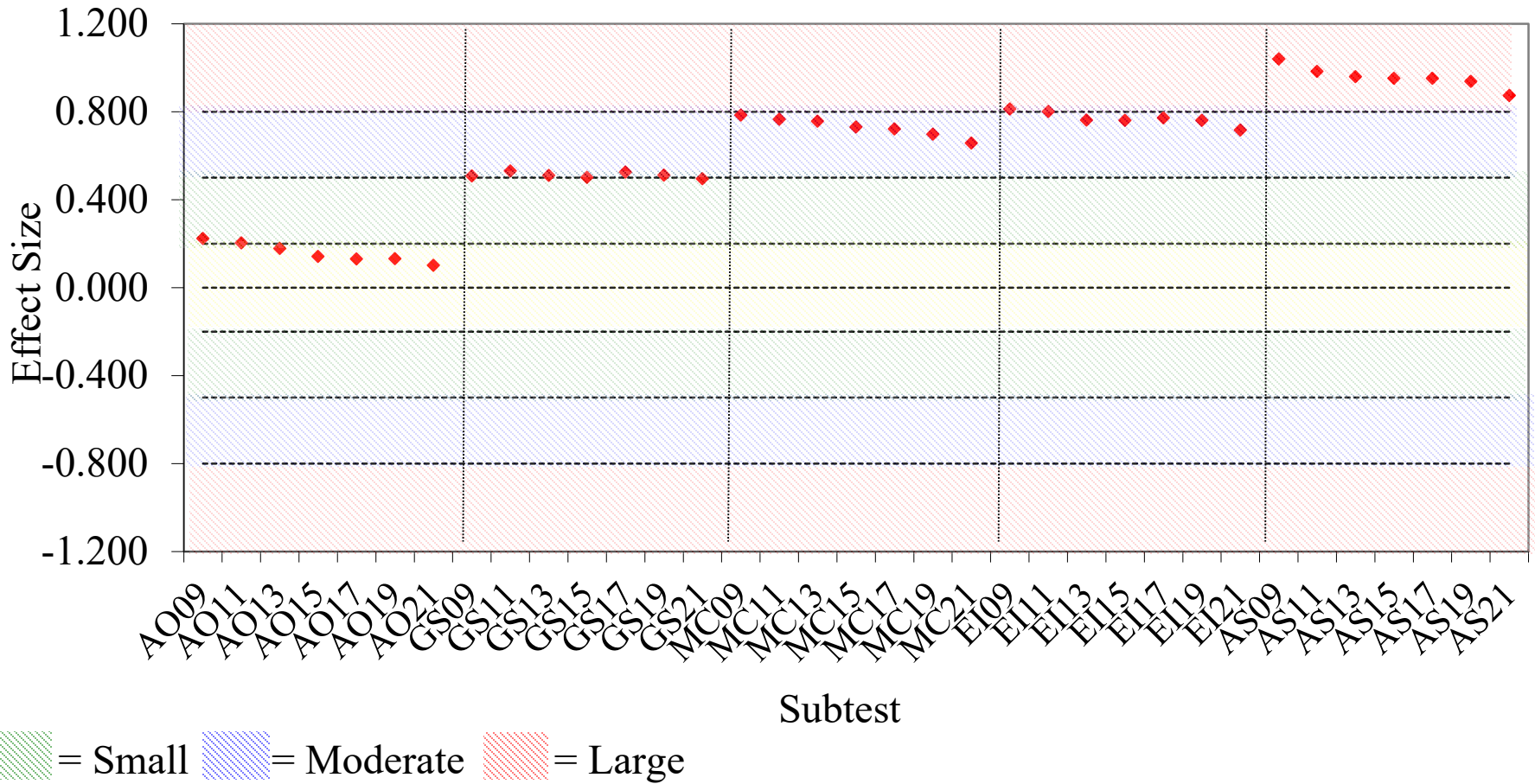
◆ IIIB  
◆ IIIA

# Comparison of Effect Sizes for Odd-Numbered FY 09-21 Males Versus Females AFQT Tests/Scores

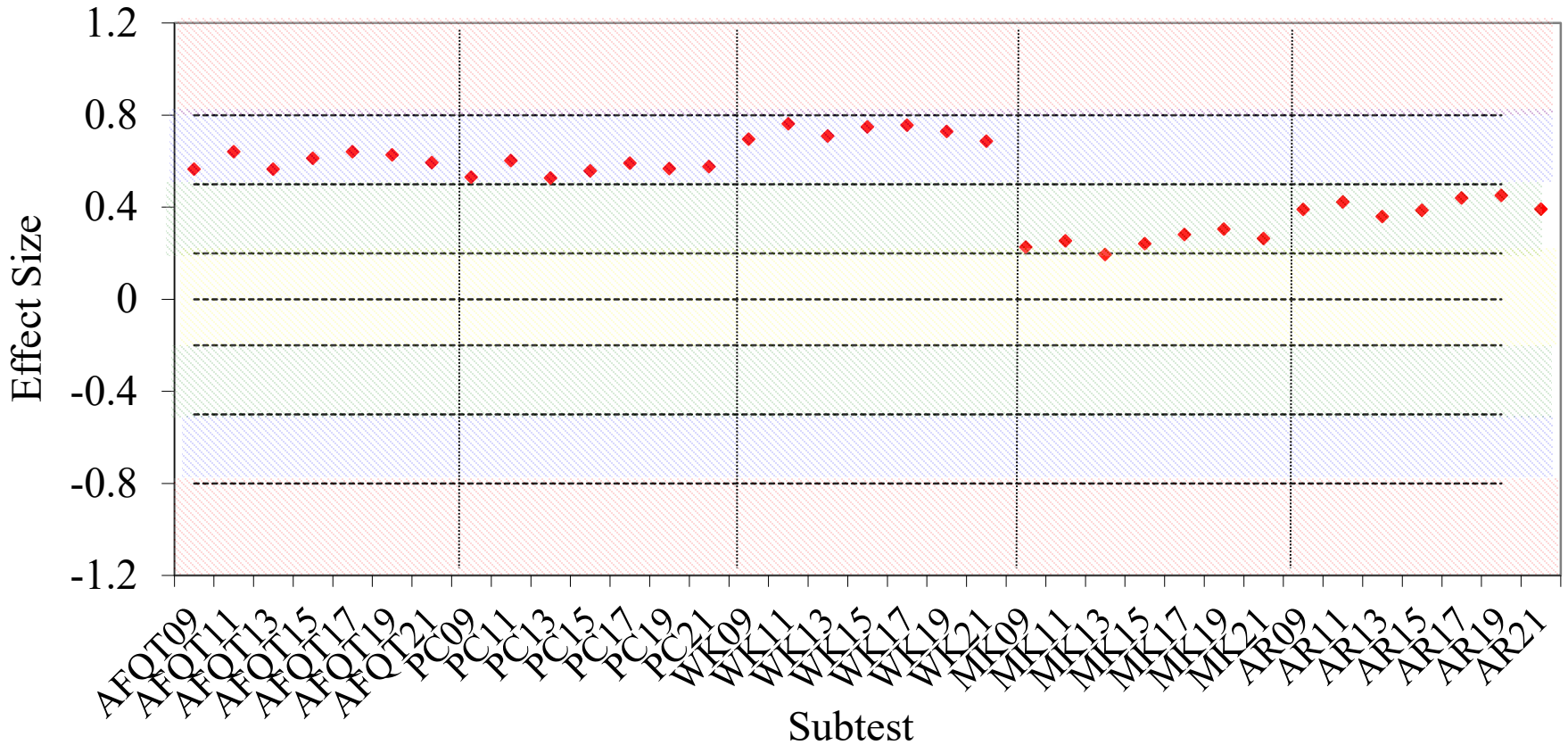


= Small
  = Moderate
  = Large

## Comparison of Effect Sizes for Odd-Numbered FYs 09-21 Males Versus Females Non-AFQT Tests

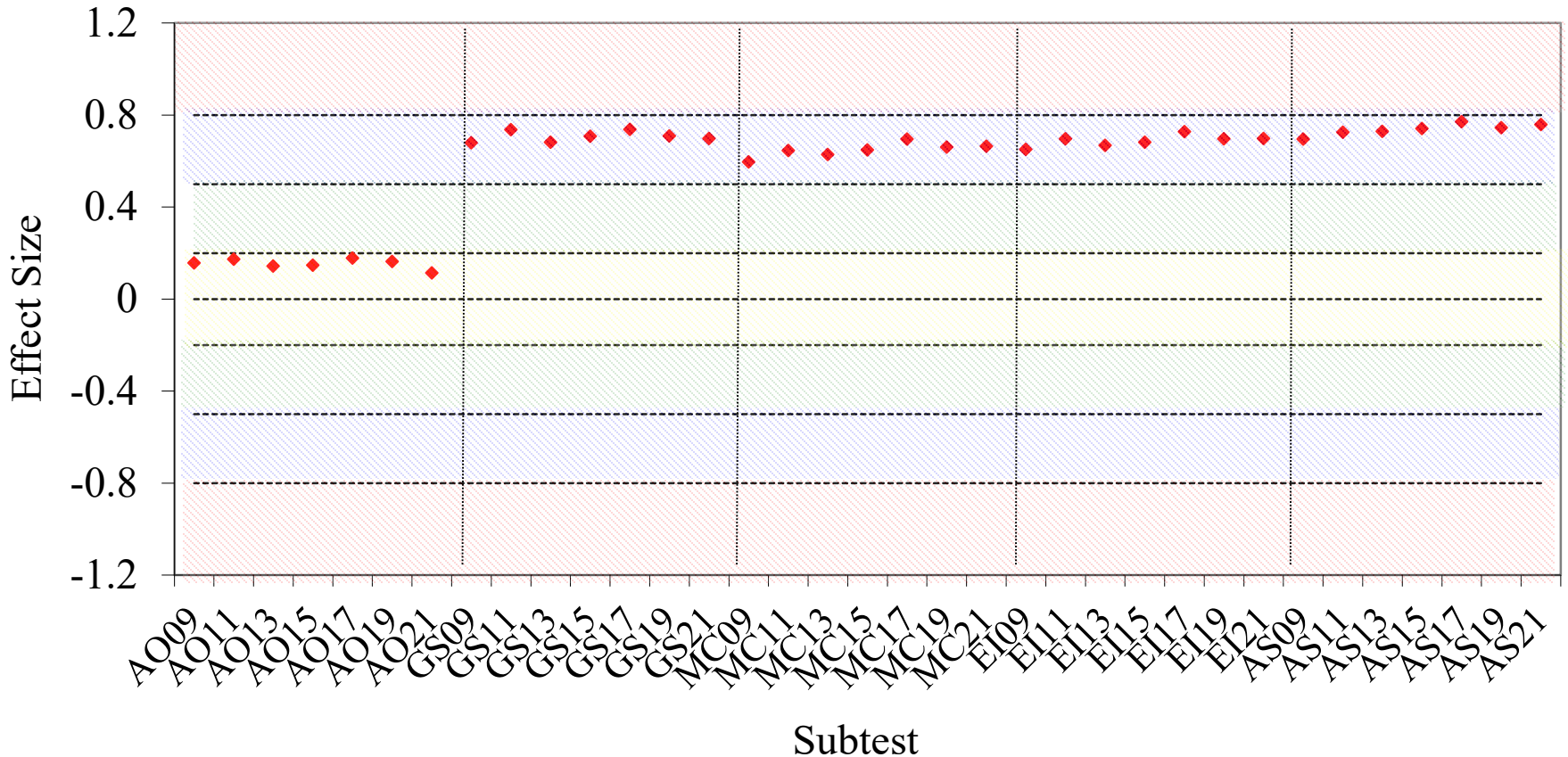


## Comparison of Effect Sizes for Odd-Numbered FYs 09-21 Non-Hispanic Whites Versus Hispanic Whites AFQT Tests/Scores



= Small
  = Moderate
  = Large

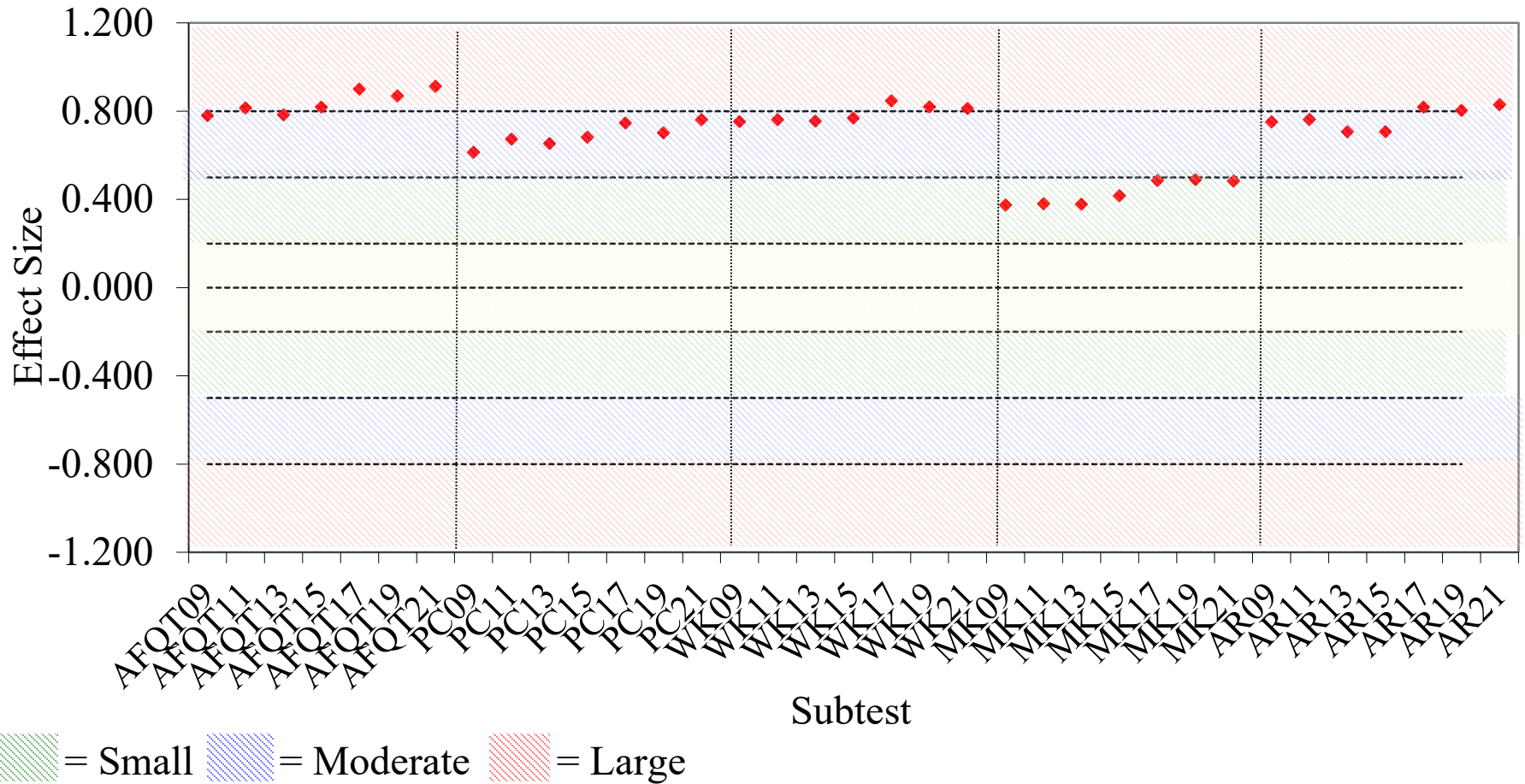
## Comparison of Effect Sizes for Odd-Numbered FYs 09-21 Non-Hispanic Whites Versus Hispanic Whites Non-AFQT Tests



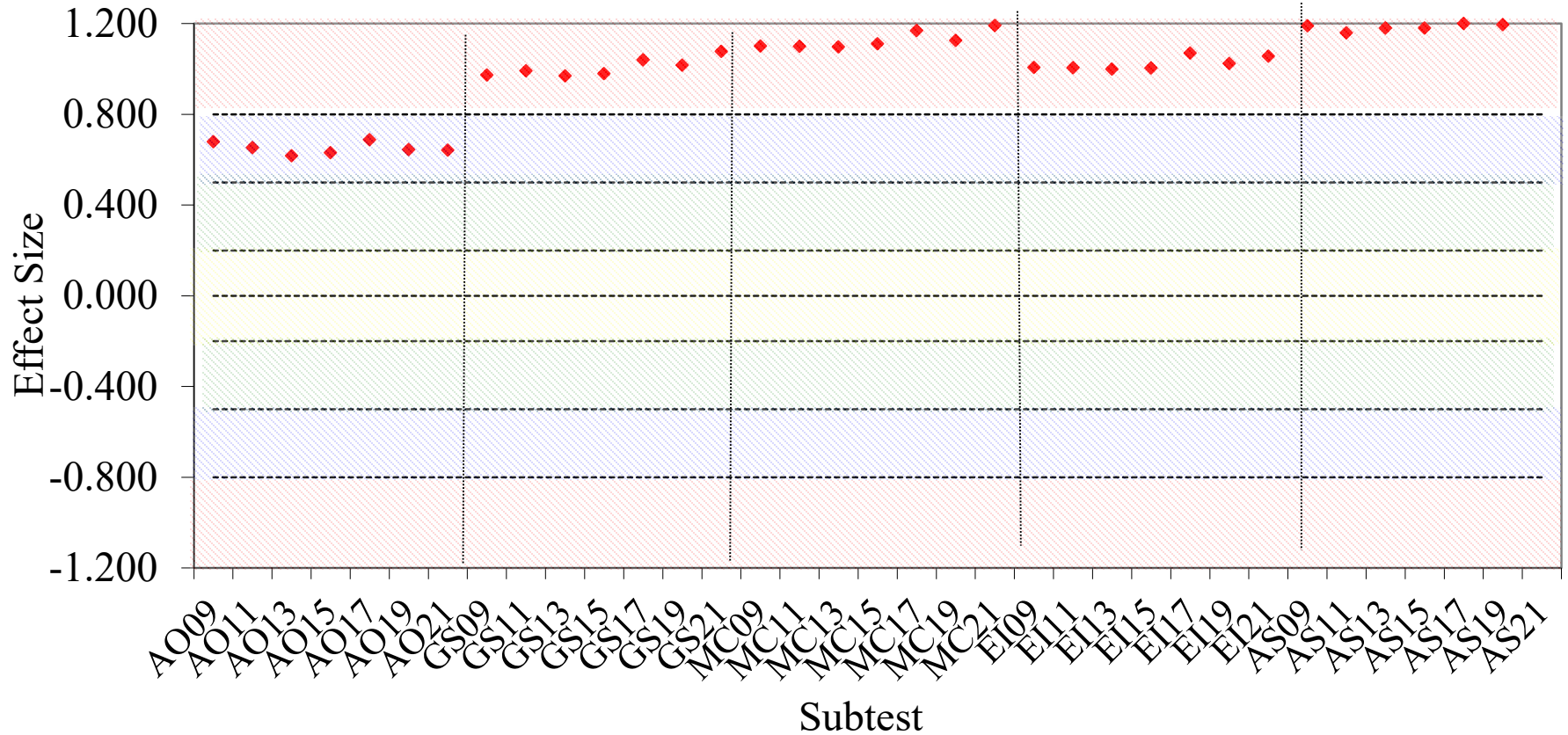
= Small
  = Moderate
  = Large



## Comparison of Effect Sizes for Odd-Numbered FYs 09-21 Non-Hispanic Whites Versus Non-Hispanic Blacks AFQT Tests/Scores

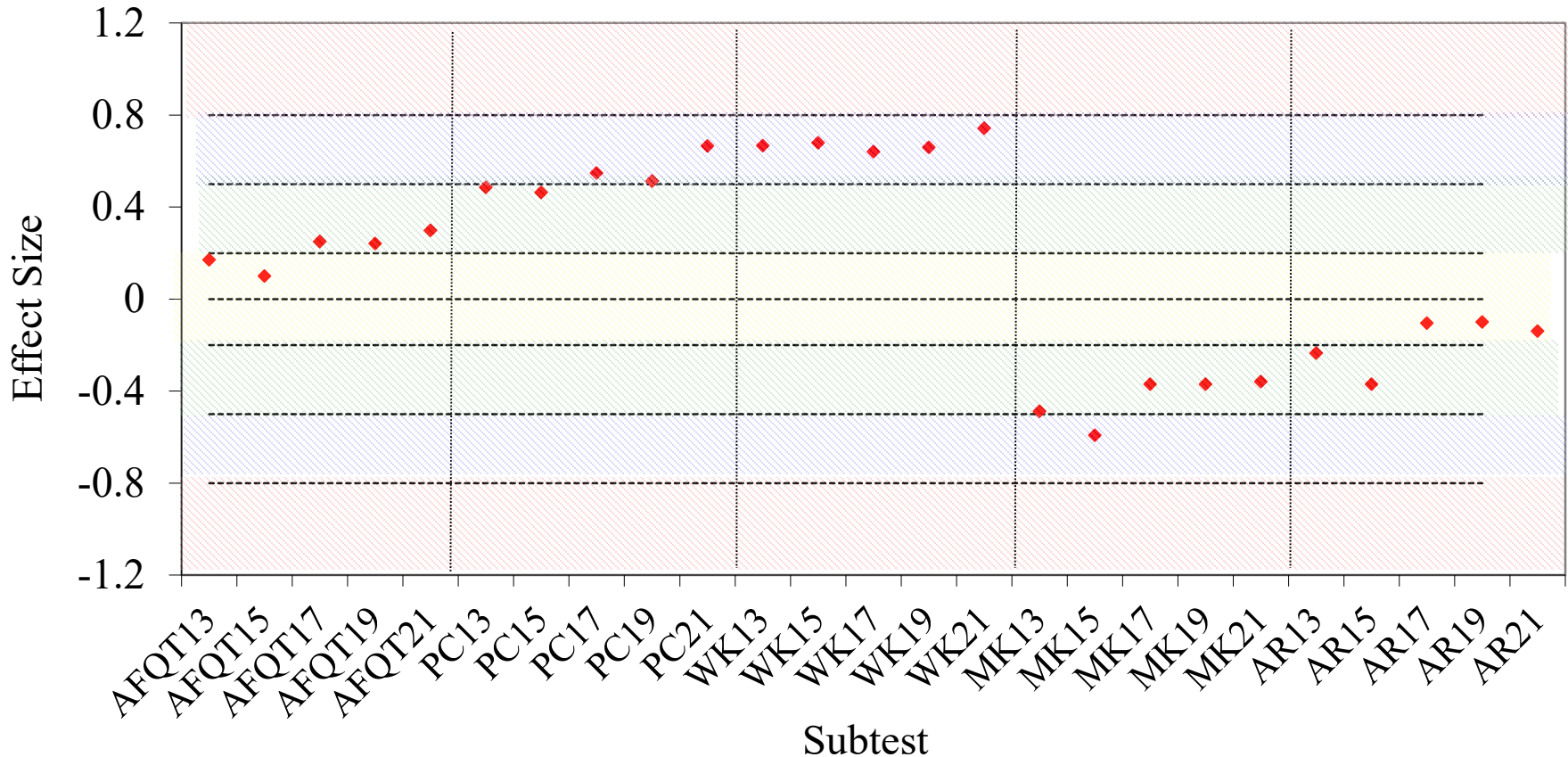


## Comparison of Effect Sizes for Odd-Numbered FYs 09-21 Non-Hispanic Whites Versus Non-Hispanic Blacks Non-AFQT Tests



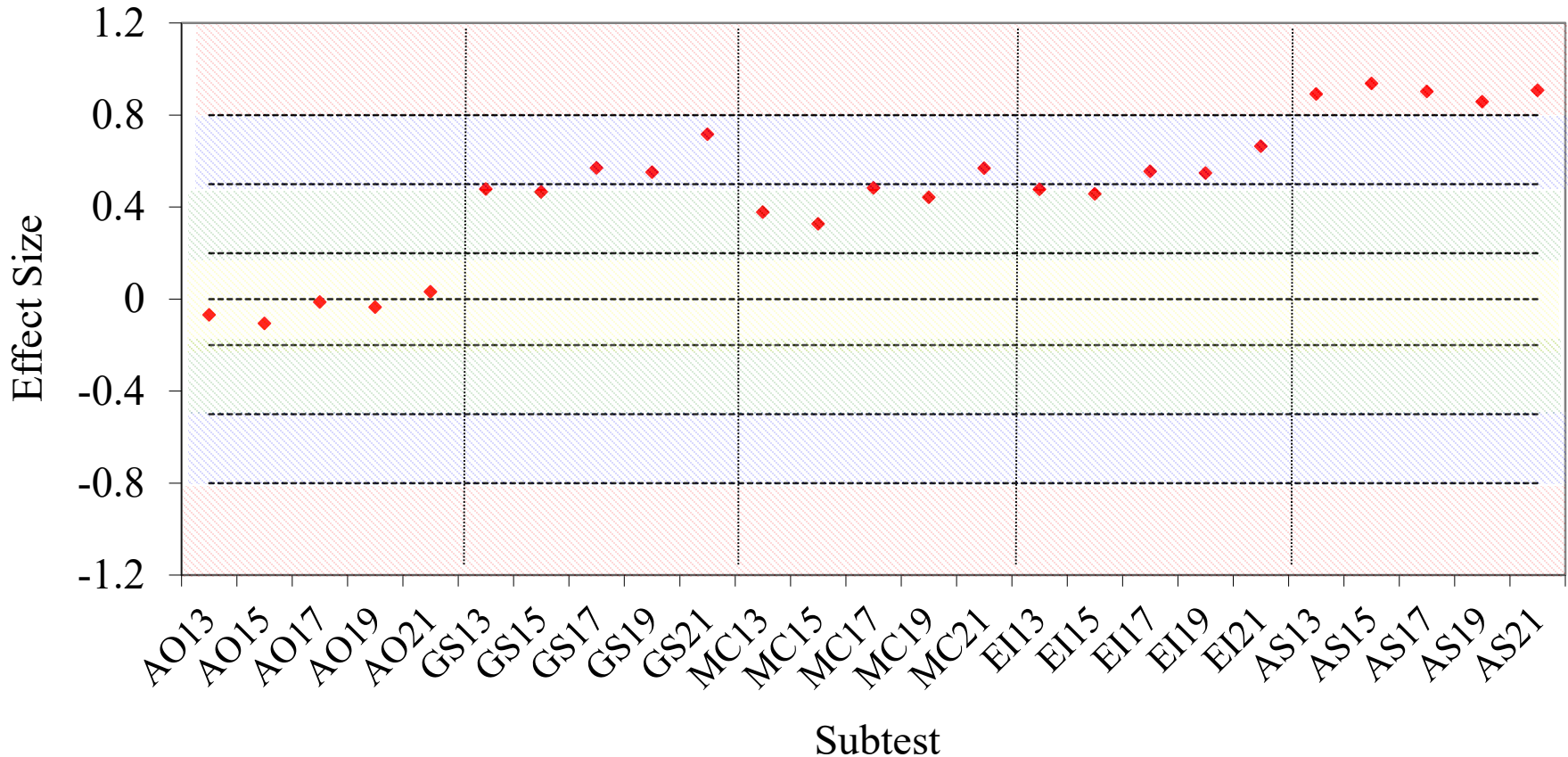
= Small
  = Moderate
  = Large

## Comparison of Effect Sizes for Odd-Numbered FYs 13-21 Non-Hispanic Whites Versus Non-Hispanic Asians AFQT Tests/Scores



= Small
  = Moderate
  = Large

## Comparison of Effect Sizes for Odd-Numbered FYs 13-21 Non-Hispanic Whites Versus Non-Hispanic Asians Non-AFQT Tests



= Small
  = Moderate
  = Large

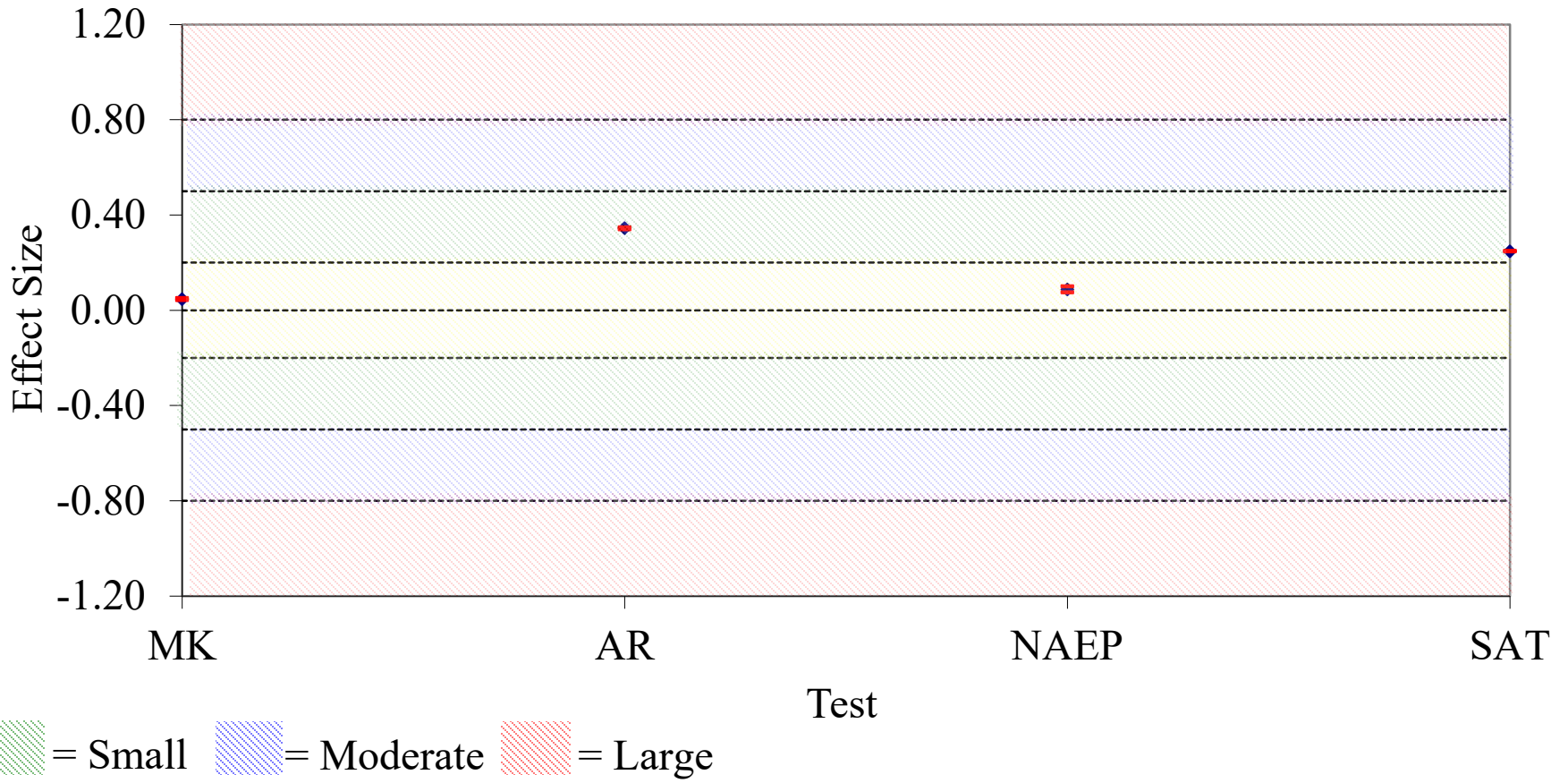
# WHAT DOES IT MEAN?




- **The magnitude of impact on the ASVAB has remained fairly constant across fiscal years, but still varies in size from negligible to large across tests and groups.**
- **A comparison of impact across different testing programs gives some indication of whether the observed FY2021 magnitudes are reasonable.**
- **Sufficient information for estimating effect sizes is available online for two other large-scale testing programs:**
  1. SAT – 2016 College Bound Seniors (Math and Reading)
  2. NAEP – 2019 Grade 12 (Reading, Math, and Science)

# Comparison of Effect Sizes Across Testing Programs

## Content Area = Math

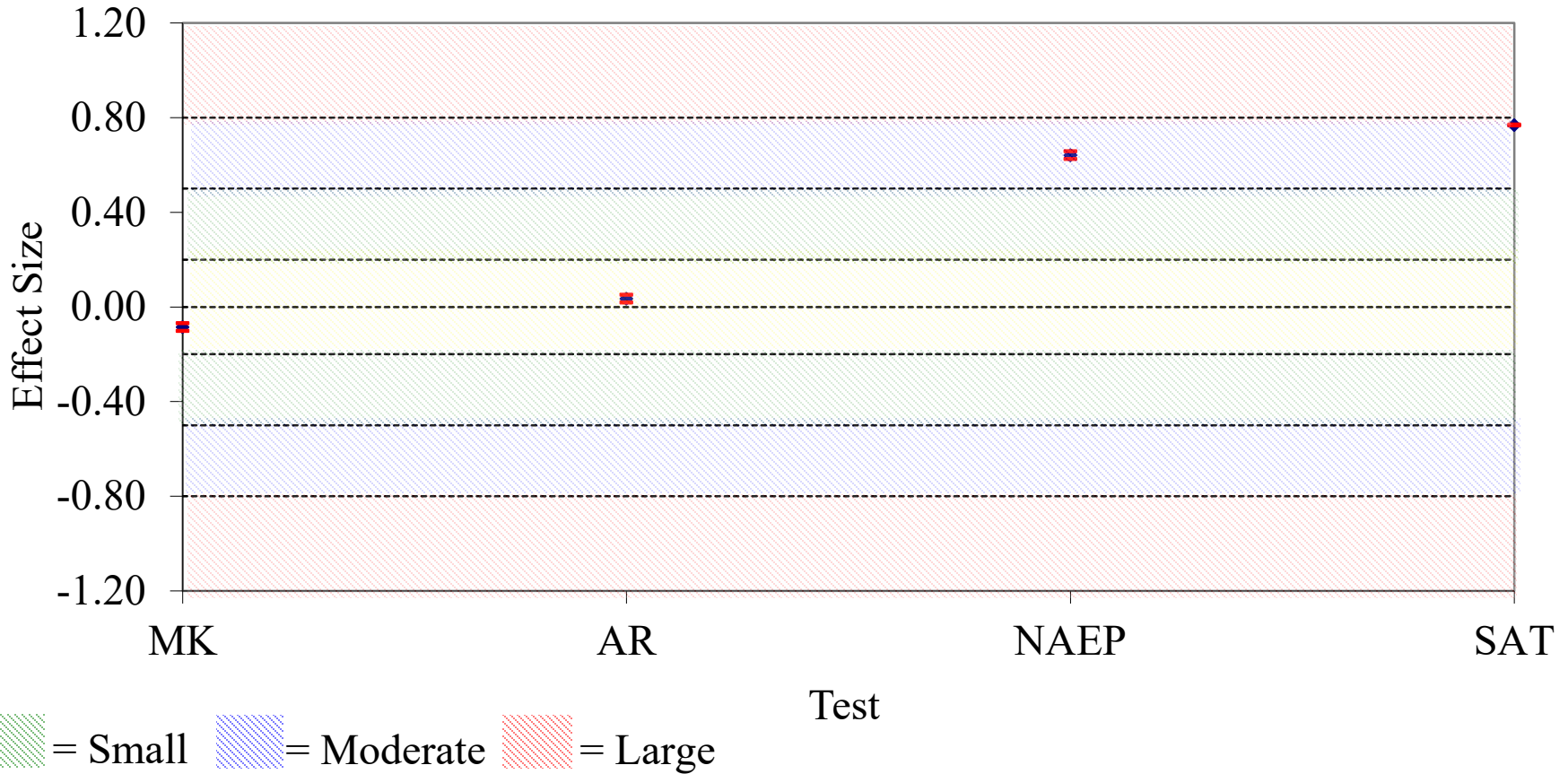
### Males Versus Females



 = Small  = Moderate  = Large

**OPA**  
OFFICE OF PEOPLE ANALYTICS

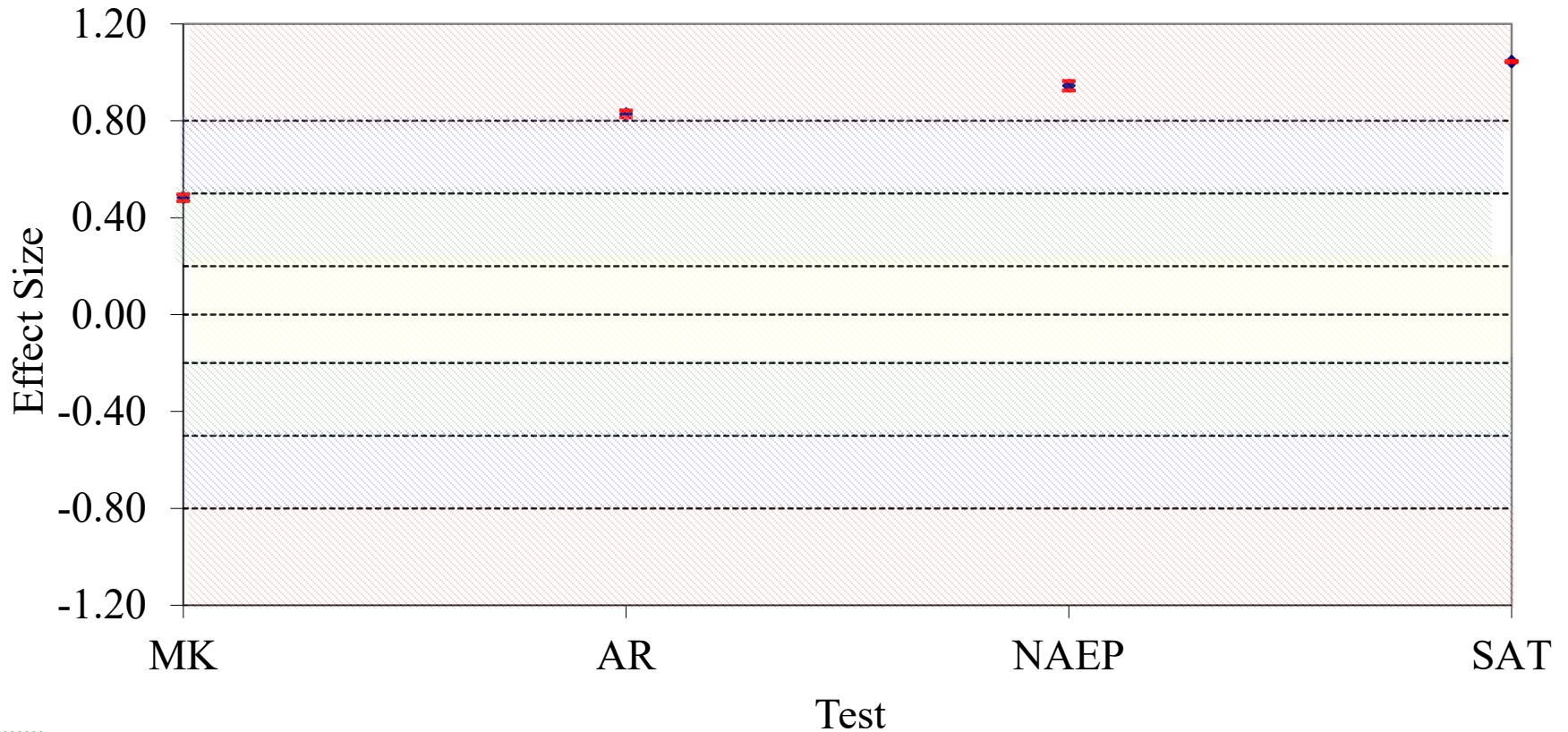
## Comparison of Effect Sizes Across Testing Programs Content Area = Math Non-Hispanic Whites Versus Hispanics\*



# Comparison of Effect Sizes Across Testing Programs

## Content Area = Math

### Non-Hispanic Whites Versus Non-Hispanic Blacks



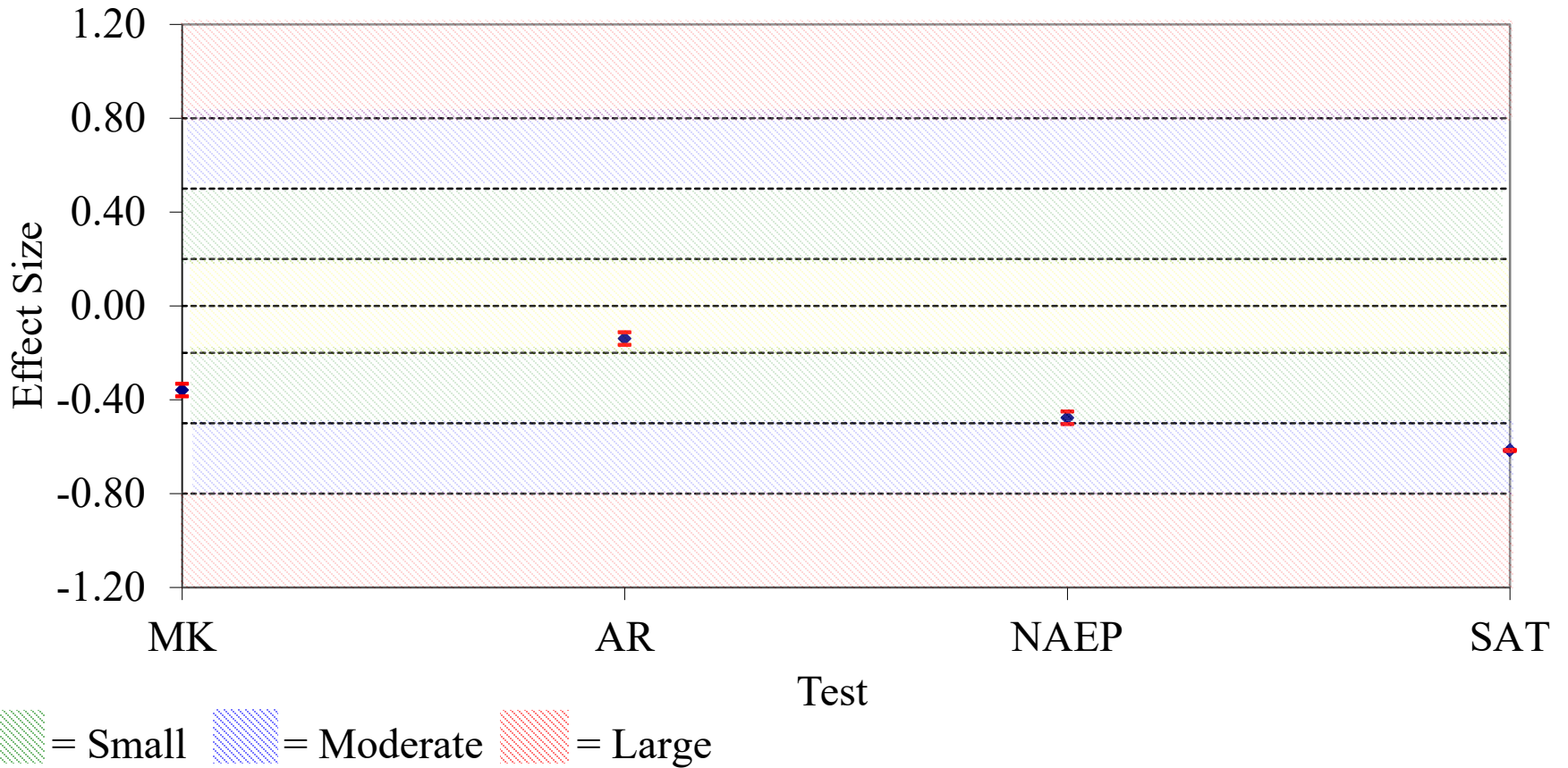
 = Small  = Moderate  = Large



# Comparison of Effect Sizes Across Testing Programs

## Content Area = Math

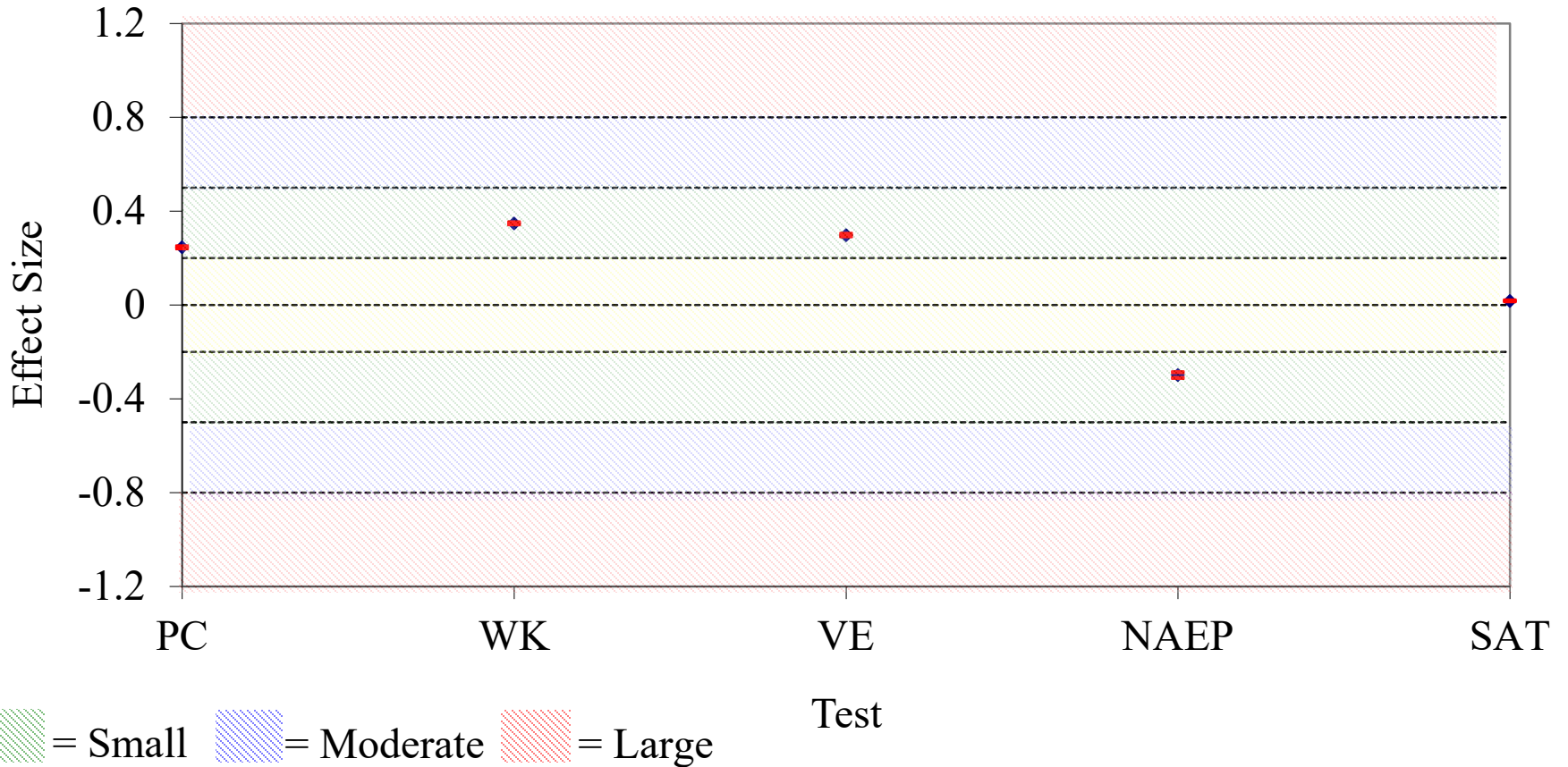
### Non-Hispanic Whites Versus Non-Hispanic Asians



# Comparison of Effect Sizes Across Testing Programs

## Content Area = Reading/Verbal

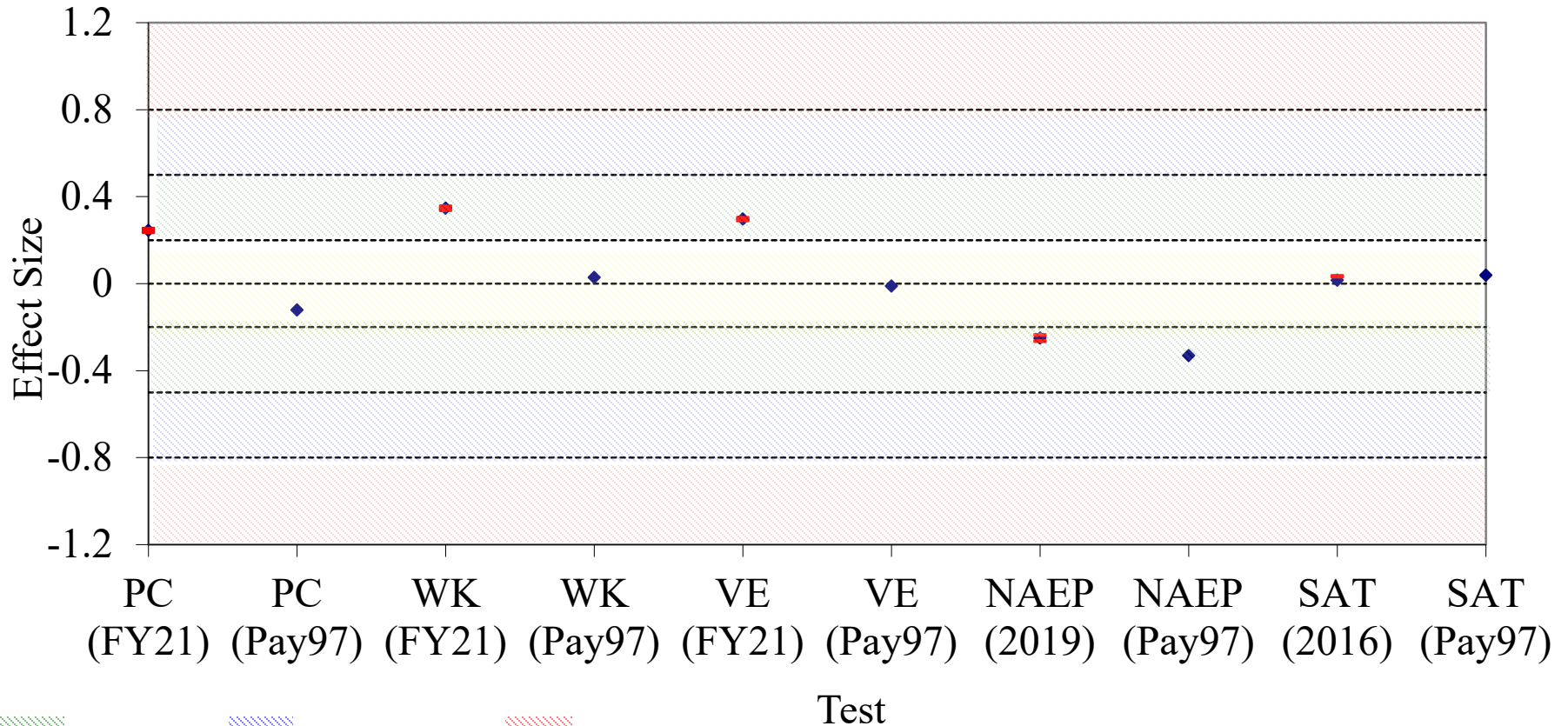
### Males Versus Females



# Comparison of Effect Sizes Across Testing Programs

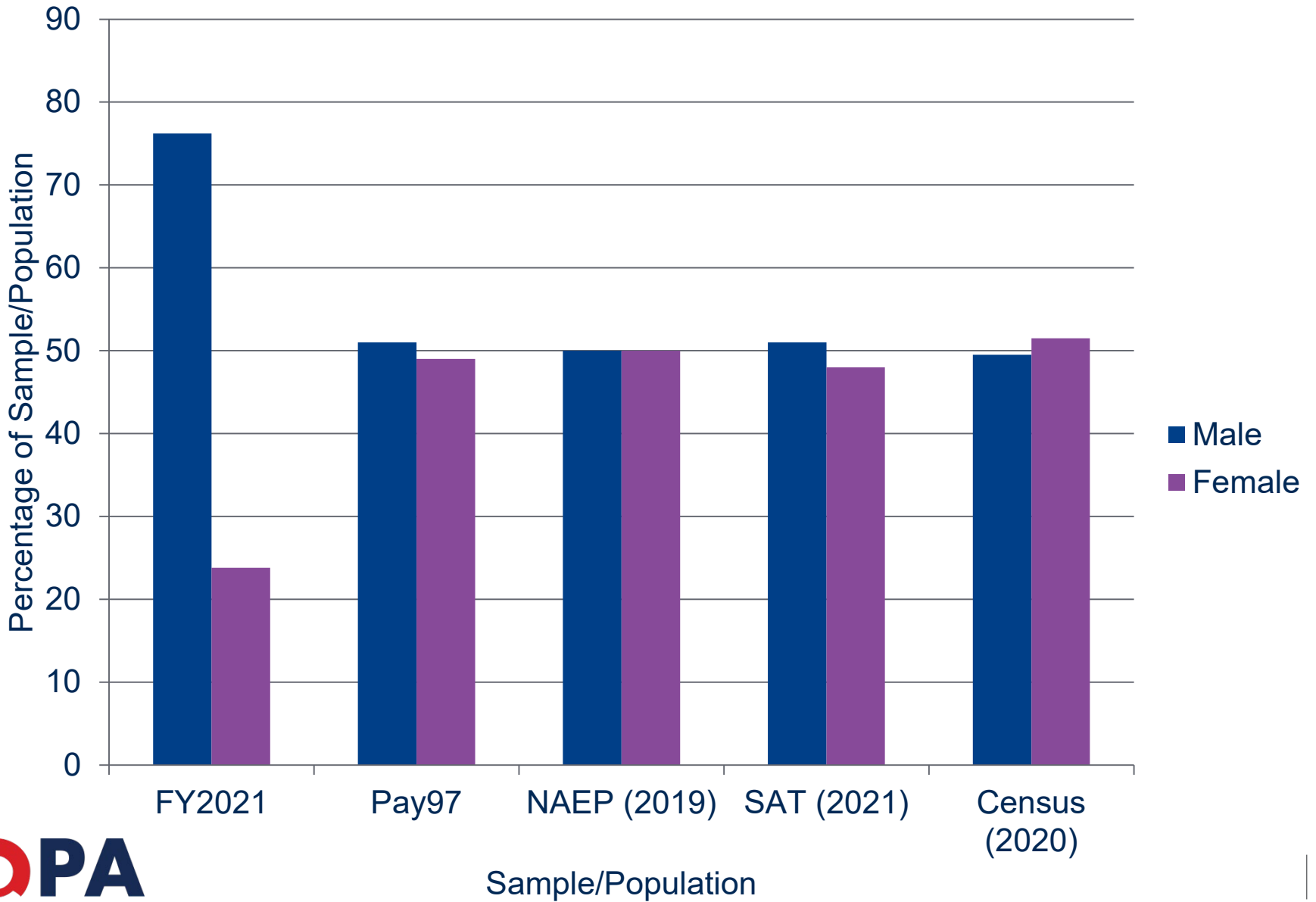
## Content Area = Reading/Verbal

### Males Versus Females



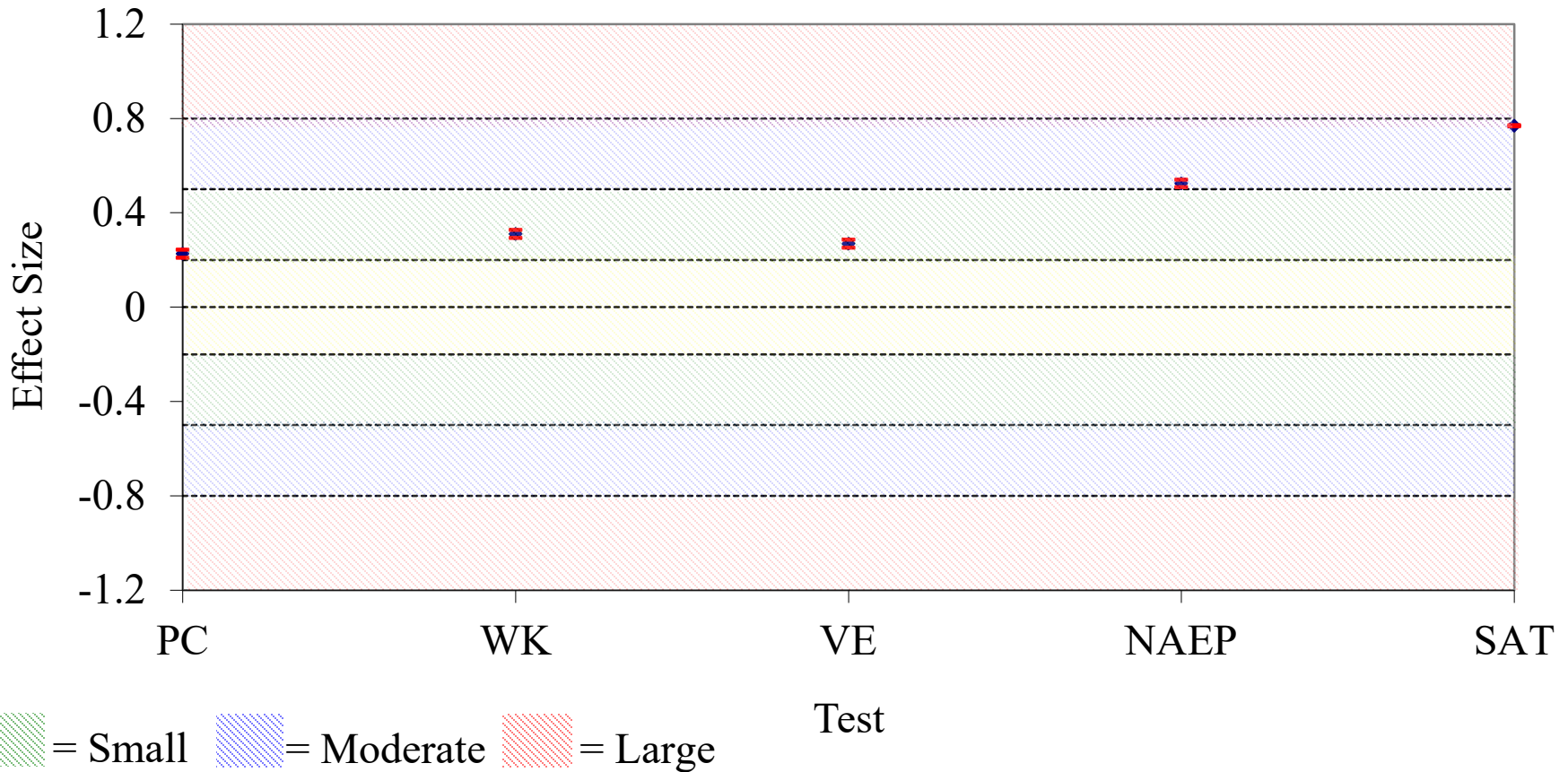
= Small
  = Moderate
  = Large

# Gender Representation Across Samples/Populations



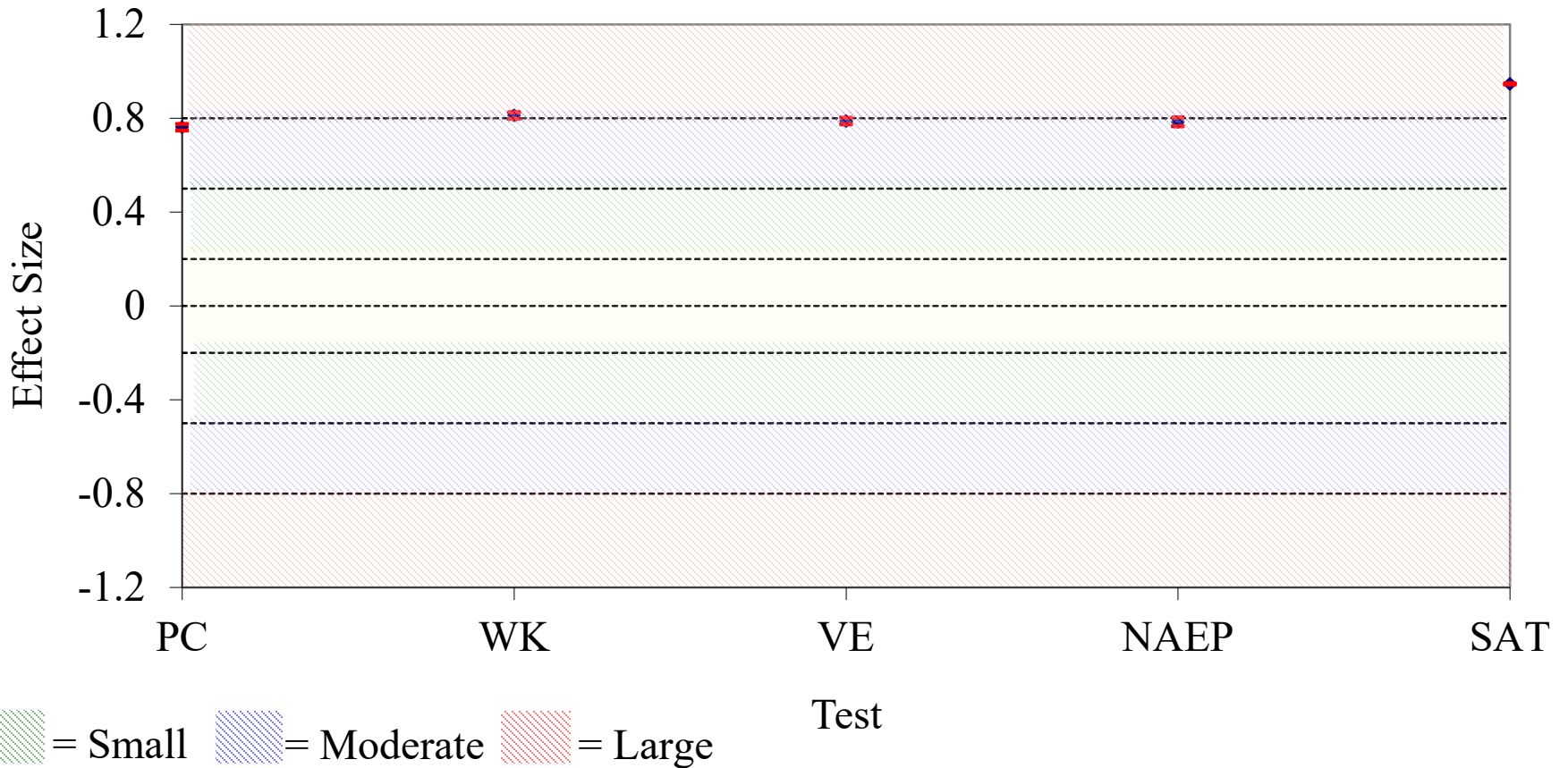
# Comparison of Effect Sizes Across Testing Programs

Content Area = Reading/Verbal  
Non-Hispanic Whites Versus Hispanics\*



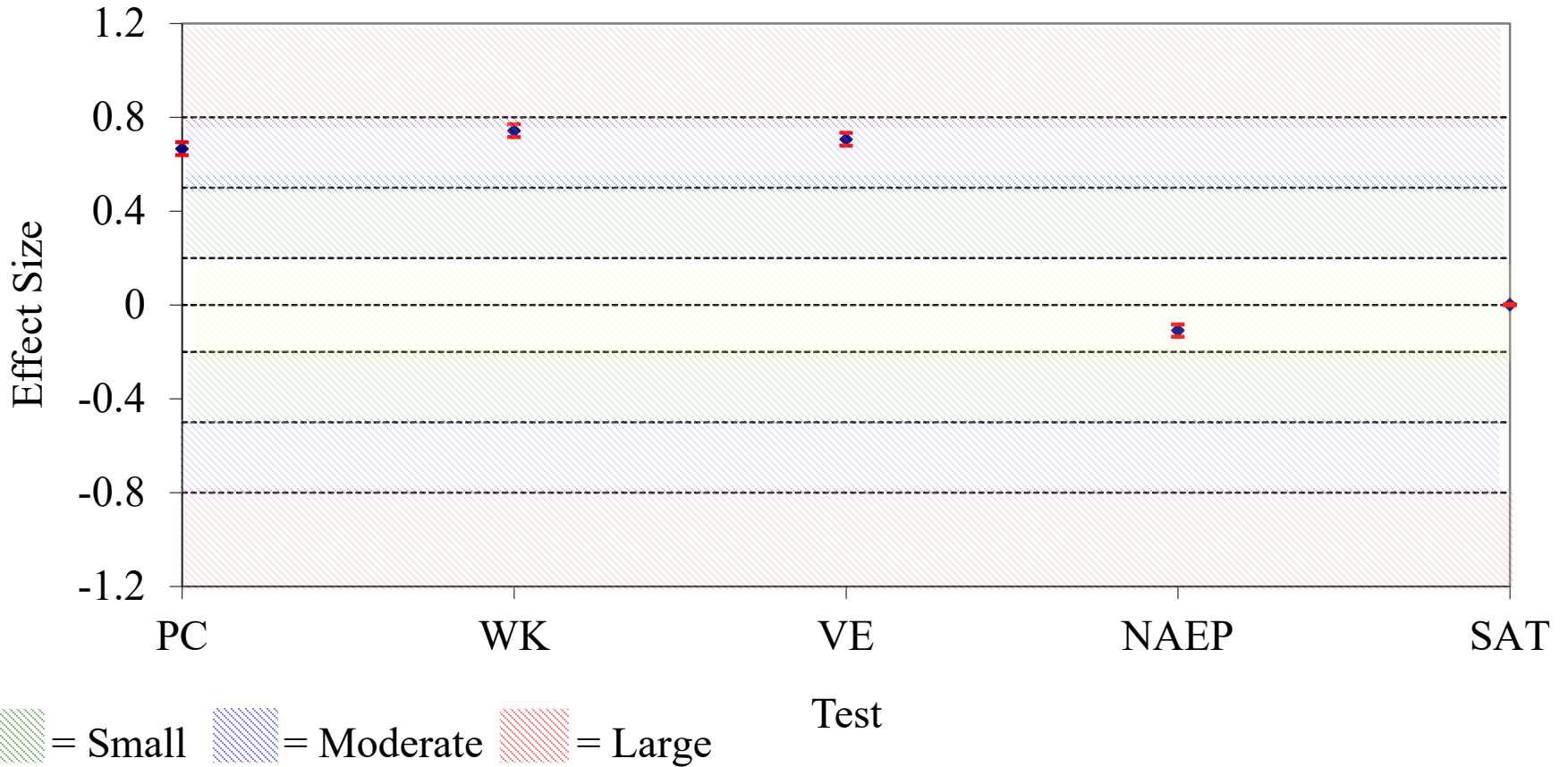
# Comparison of Effect Sizes Across Testing Programs

Content Area = Reading/Verbal  
Non-Hispanic Whites Versus Non-Hispanic Blacks



# Comparison of Effect Sizes Across Testing Programs

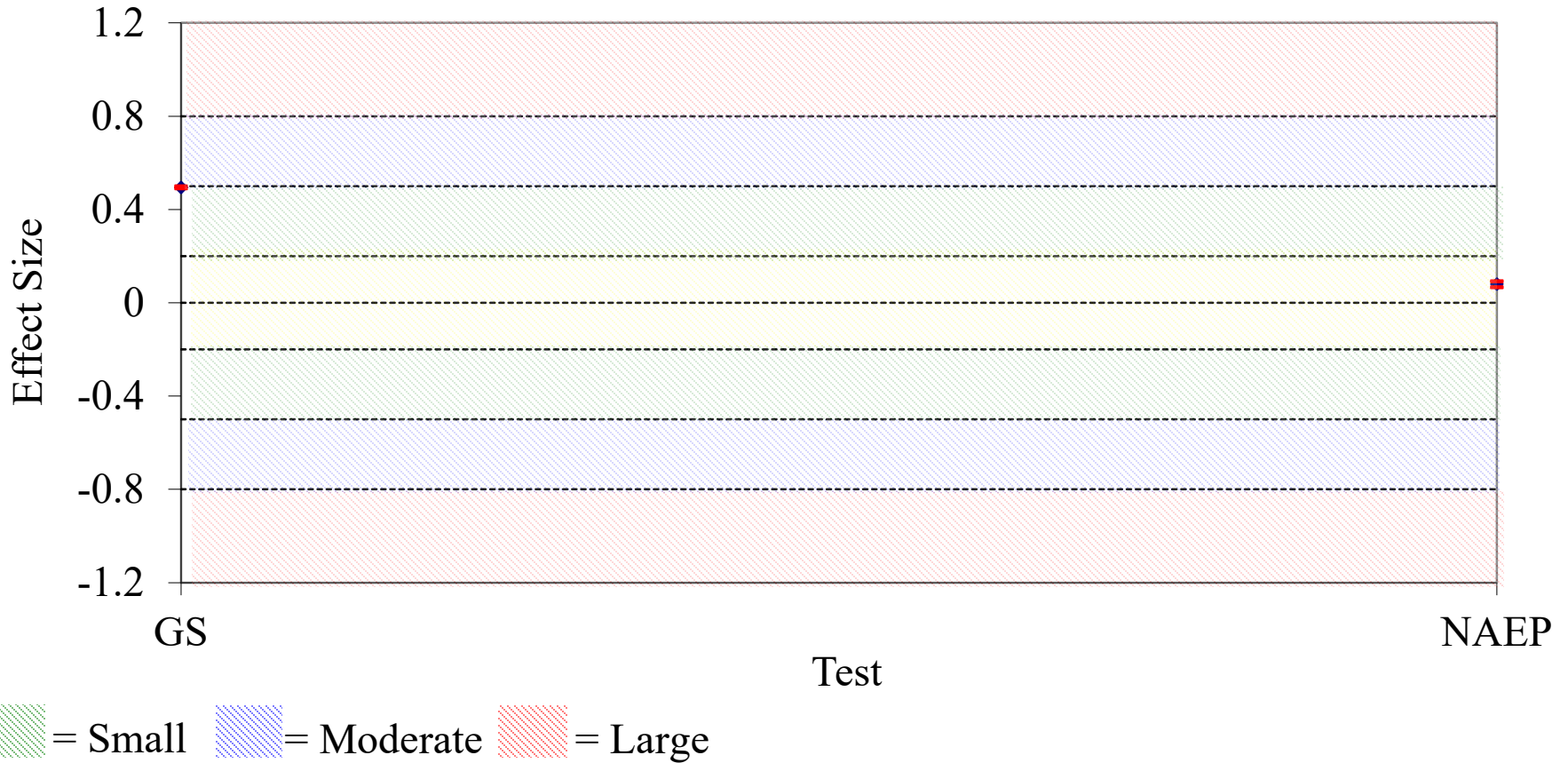
Content Area = Reading/Verbal  
Non-Hispanic Whites Versus Non-Hispanic Asians



# Comparison of Effect Sizes Across Testing Programs

## Content Area = Science

### Males Versus Females

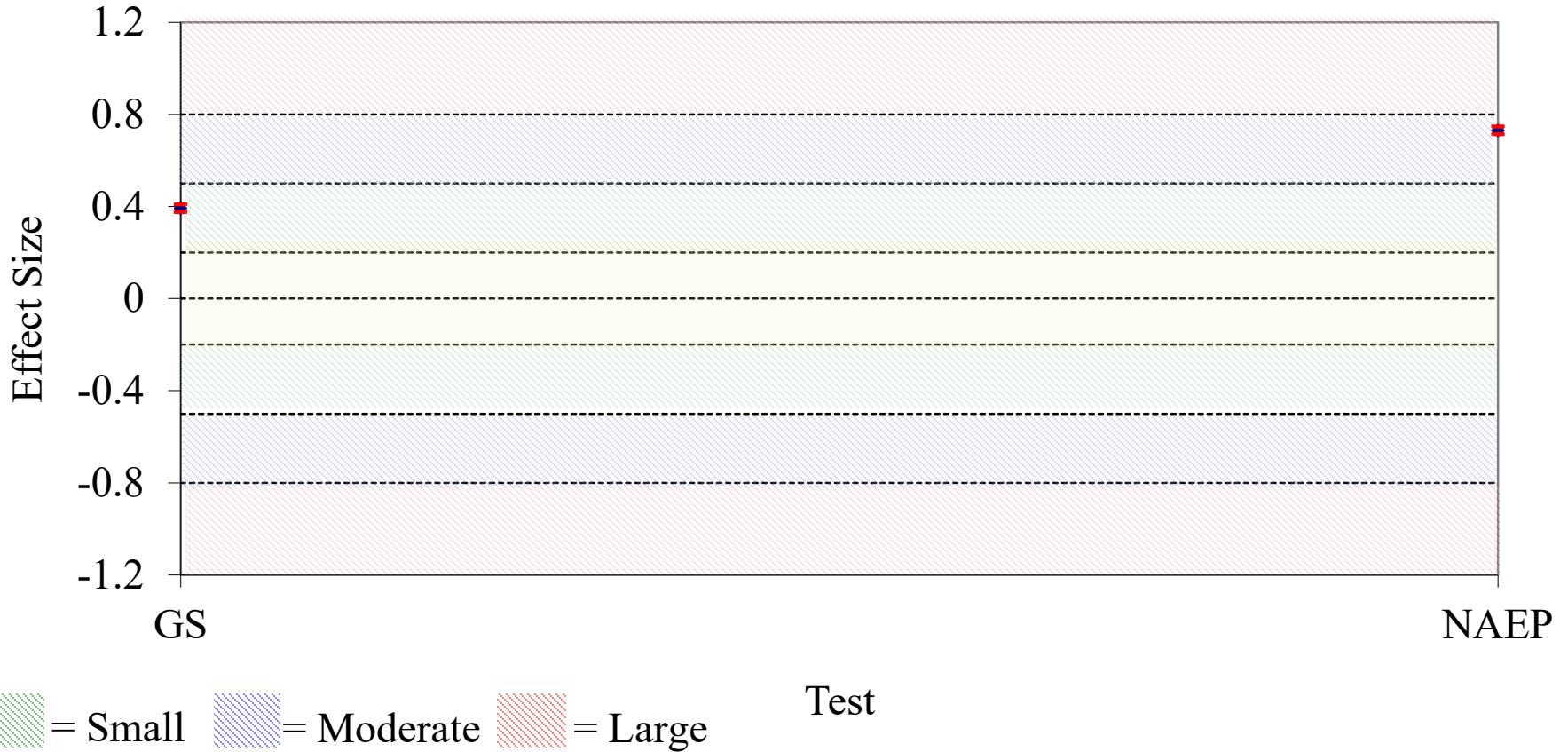




# Comparison of Effect Sizes Across Testing Programs

## Content Area = Science

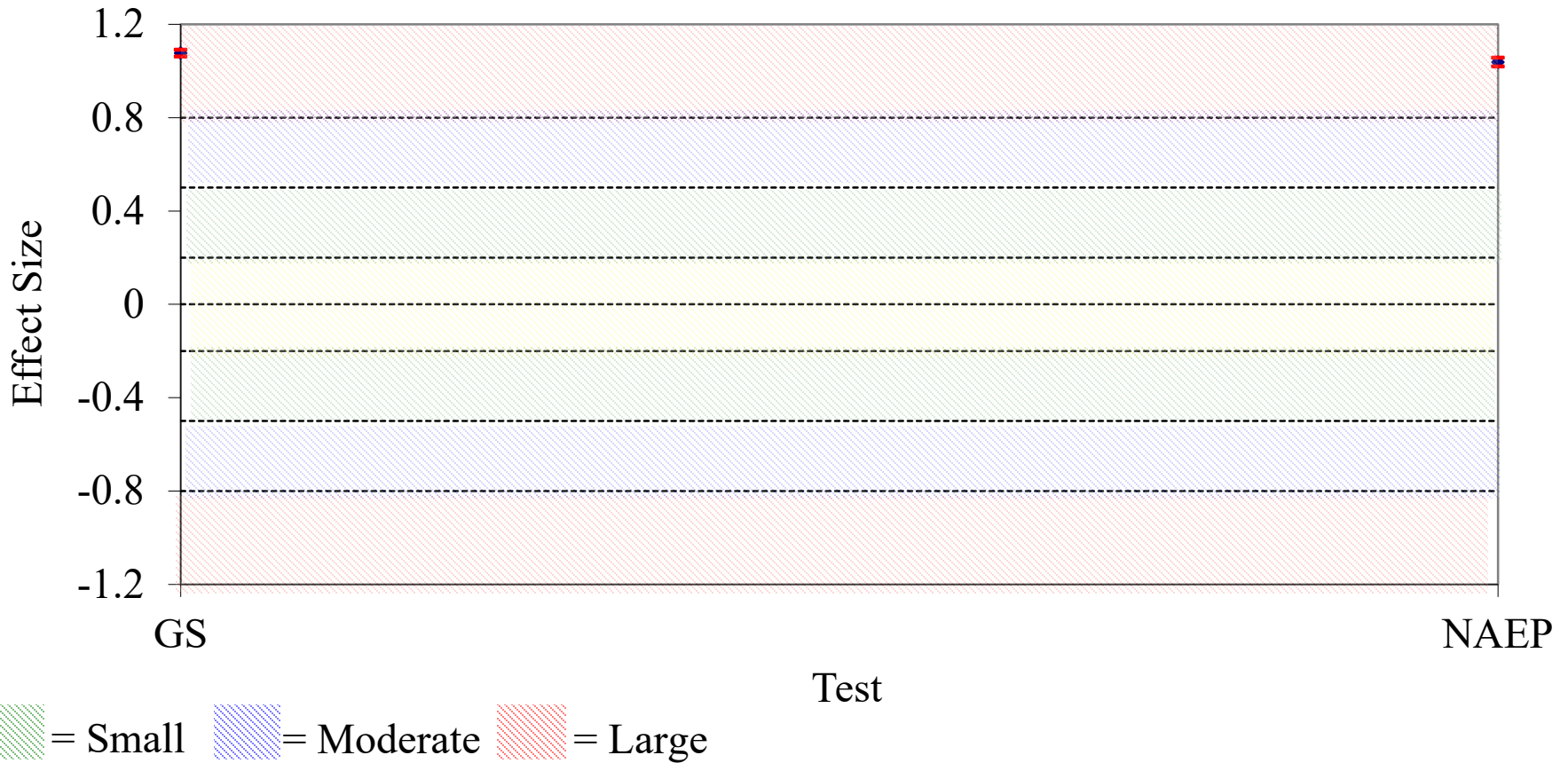
### Non-Hispanic Whites Versus Hispanics\*



# Comparison of Effect Sizes Across Testing Programs

## Content Area = Science

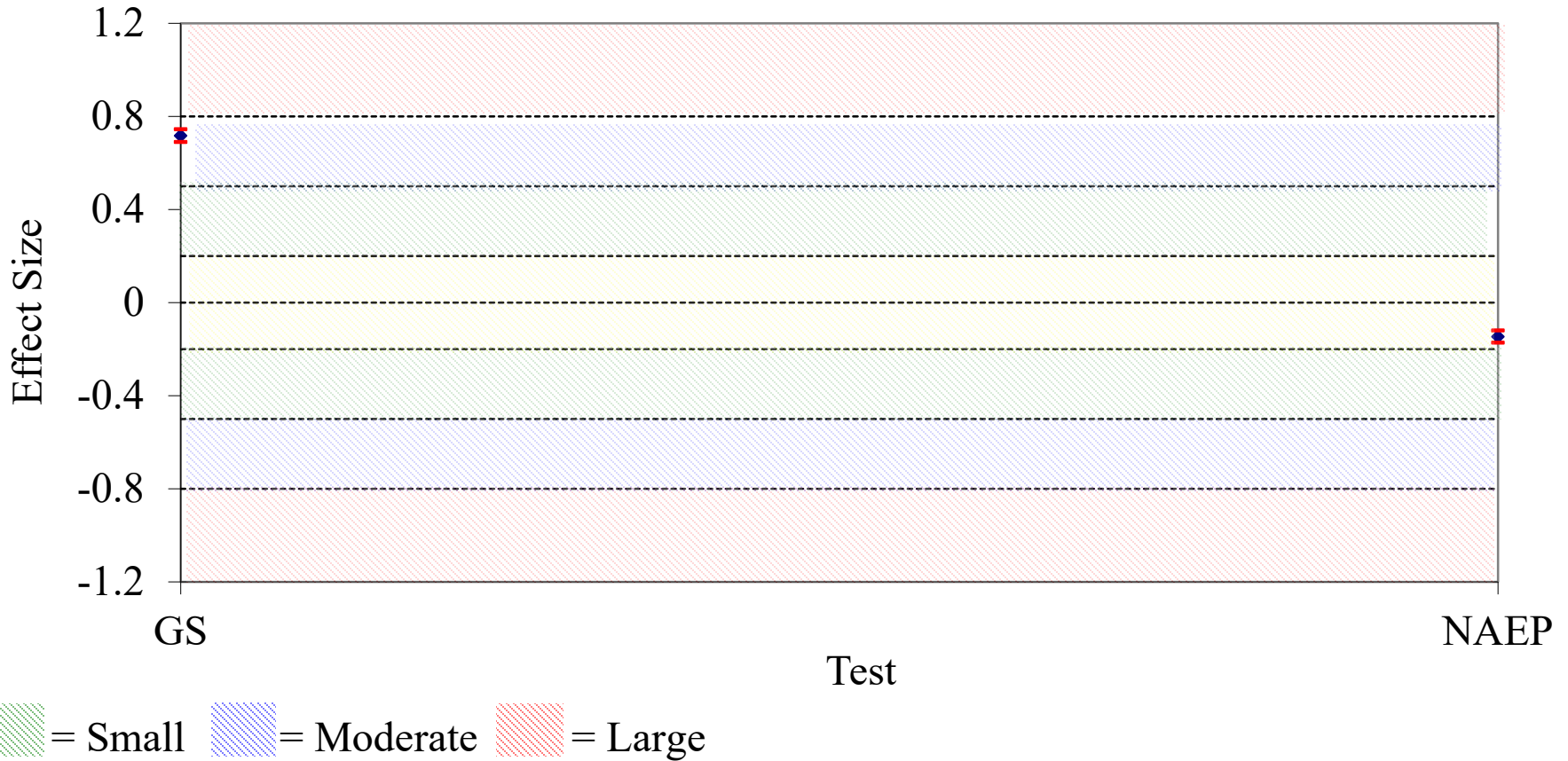
### Non-Hispanic Whites Versus Non-Hispanic Blacks



# Comparison of Effect Sizes Across Testing Programs

## Content Area = Science

### Non-Hispanic Whites Versus Non-Hispanic Asians

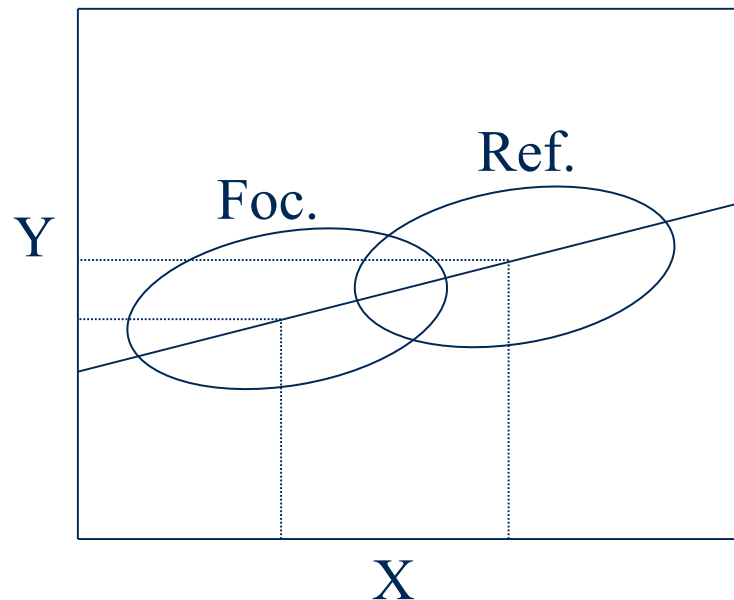


# CONCLUSIONS AND CAVEATS

- For the AFQT tests (and GS), the direction and magnitude of overall impact is generally consistent with that observed on comparable SAT and NAEP tests, which suggests that impact on ASVAB tests may reflect legitimate differences in the studied groups.
  - Comparisons across programs may be somewhat restricted due to differences in group definitions, testing populations, test content, etc.
- “To the extent that members of one group do more poorly on a subtest of items that are a *legitimate part of the content domain*, we would be reluctant to call the discrepancy evidence of *bias*” (Shepard, 1987).

# CONCLUSIONS AND CAVEATS

- **Adverse impact does *not* reflect test bias if validity research shows that the test is equally valid for relevant groups.**
  - Historically, a regression-based approach has been advocated to evaluate the existence of bias. Lack of test bias is indicated when the regression line relating the test score [X] and a criterion [Y] is the same for each group.



# CONCLUSIONS AND CAVEATS

- **Previous research on the ASVAB technical tests showed similar prediction lines across (1) males and females and (2) blacks and whites (Wise, et al., 1992), suggesting no bias for the tests and groups studied.**
  - DMDC recommended in 2010 that an updated validity study be conducted for relevant tests and groups.
- ***What's new?* Acquisition of training outcome from the Services has made it possible to examine the AFQT for differential prediction (test bias).**
- ***Now completed:* The largest military-sample differential prediction study conducted to date. See next presentation: Putka et al., (2022).**

# SPECIAL TESTS ON ASVAB PLATFORM

- **Cyber Test (Cyber):** Test of basic computer and information systems knowledge (All Services)
- **Coding Speed (CS):** A speeded test of assigning code numbers to words (Navy only)

# CONCLUSIONS FOR SPECIAL TESTS

- Cyber Test and Coding Speed generally exhibited small to moderate effects and were usually as low or lower than most ASVAB tests.
- Coding Speed usually had very small effects (near 0), BUT, this test may suffer from other issues. Some examples (also see backup slide for full list):
  - Affected by lag time in internet delivery (speeded test)
  - Delivery device and context may affect responses
  - Suffers from coachability and susceptibility to invalid strategies that result in high scores
- Potential for adverse impact is not the only consideration for making changes to the ASVAB.



# BACKUP SLIDES

# HOW IS ADVERSE IMPACT ASSESSED?

- **Statistical significance of the impact ratio can be computed, as well as confidence intervals around the impact ratio (Morris & Lobsenz, 2000):**

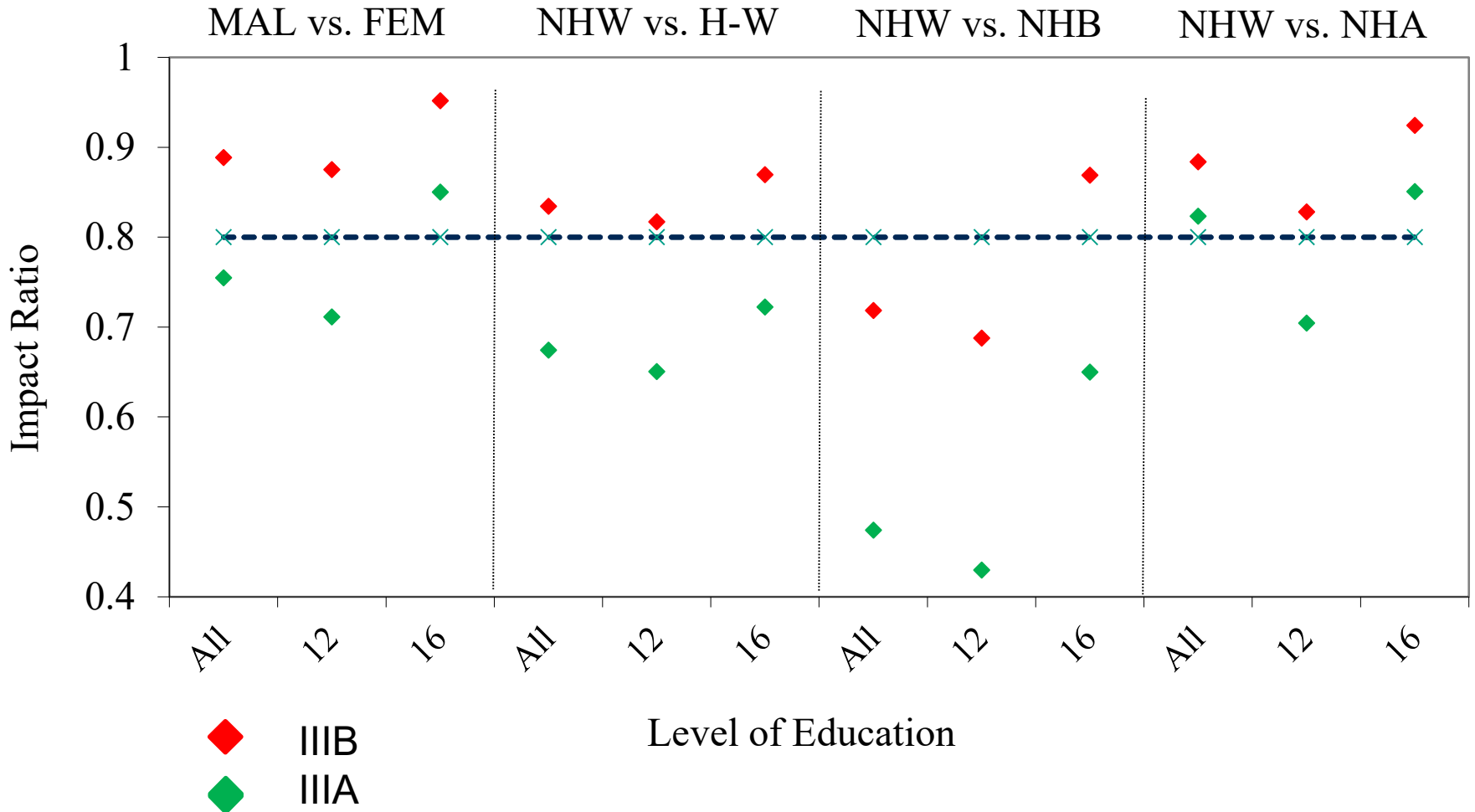
- $Z_{IR} = \frac{\ln\left(\frac{SR_{Foc}}{SR_{Ref}}\right)}{\sqrt{\frac{1-SR_{Tot}}{SR_{Tot}}\left(\frac{1}{N_{Foc}} + \frac{1}{N_{Ref}}\right)}}$ , **where  $SR$  = selection rate**

- $Z_{IR}$  is significant at  $\alpha = .05$  if  $|Z| > 1.96$

- Confidence interval =  $e^{(\ln(IR) \pm 1.96SE_{IR})}$ , **where**

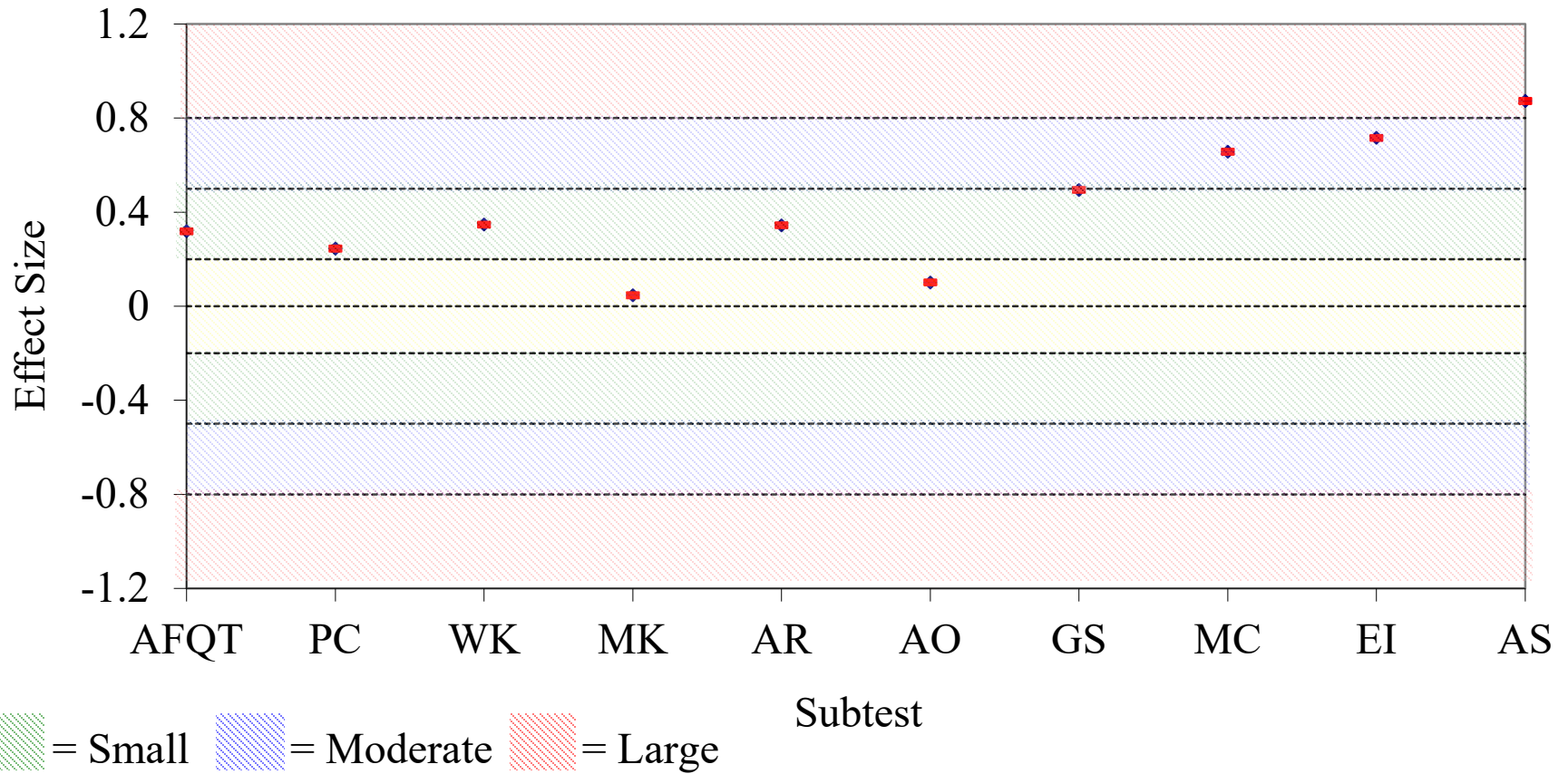
- $SE_{IR} = \sqrt{\frac{1-SR_{Foc}}{N_{Foc}SR_{Foc}} + \frac{1-SR_{Ref}}{N_{Ref}SR_{Ref}}}$

## Comparison of FY2021 Impact Ratios for Years of Education Group

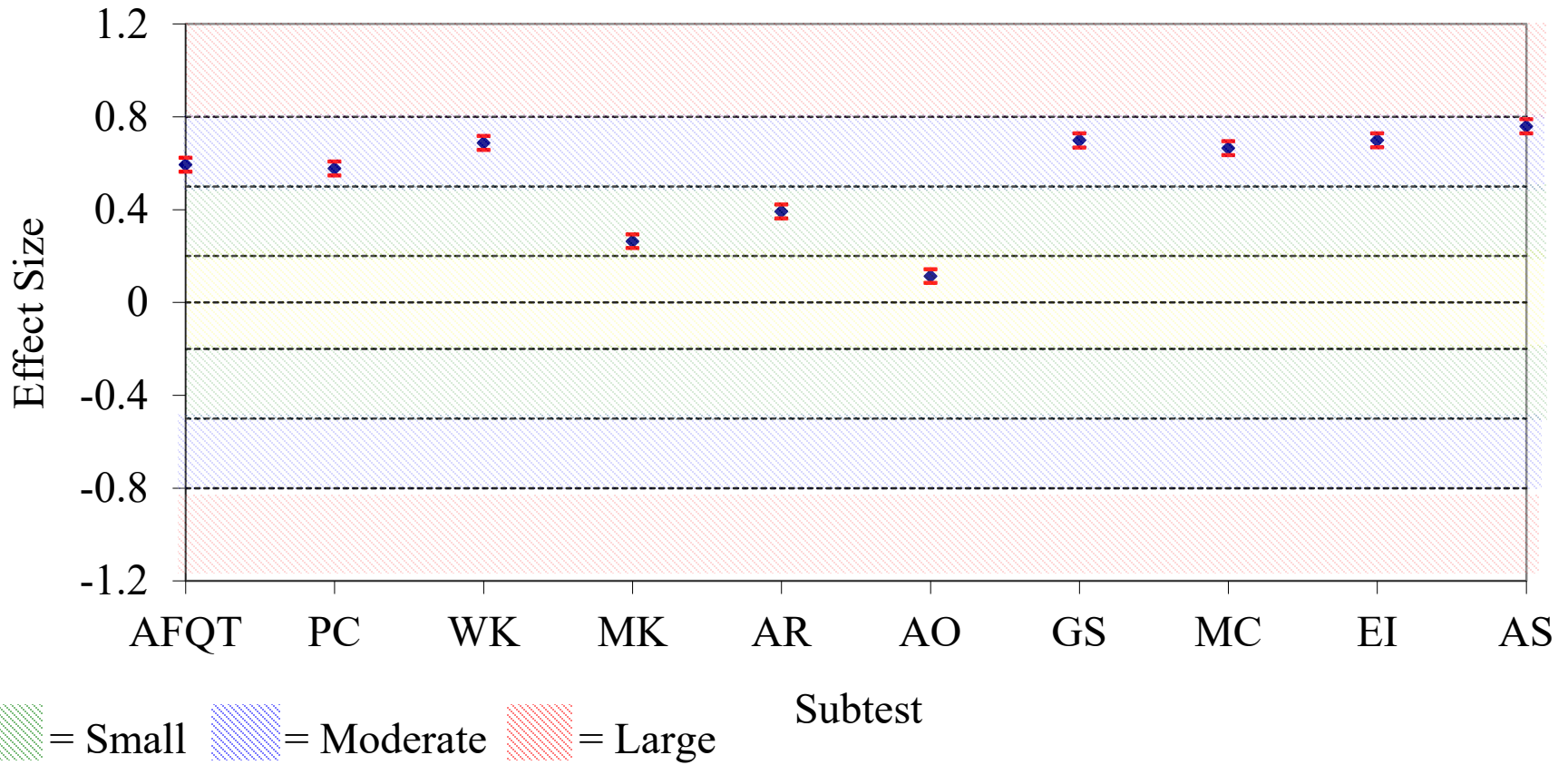


\*12 = 12 years of education reported; 16 = 16 years of education reported.

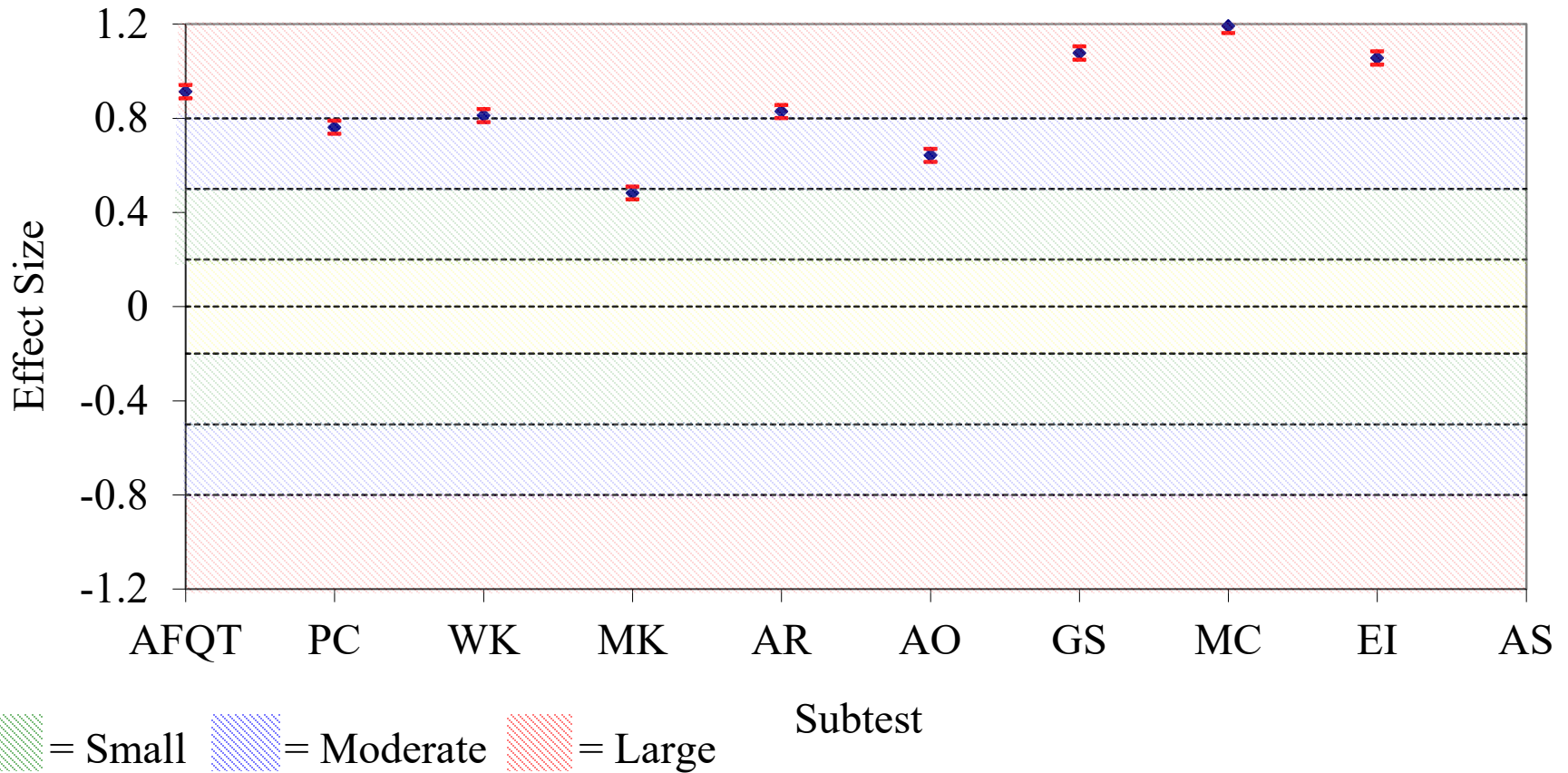
## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Males Versus Females FY2021



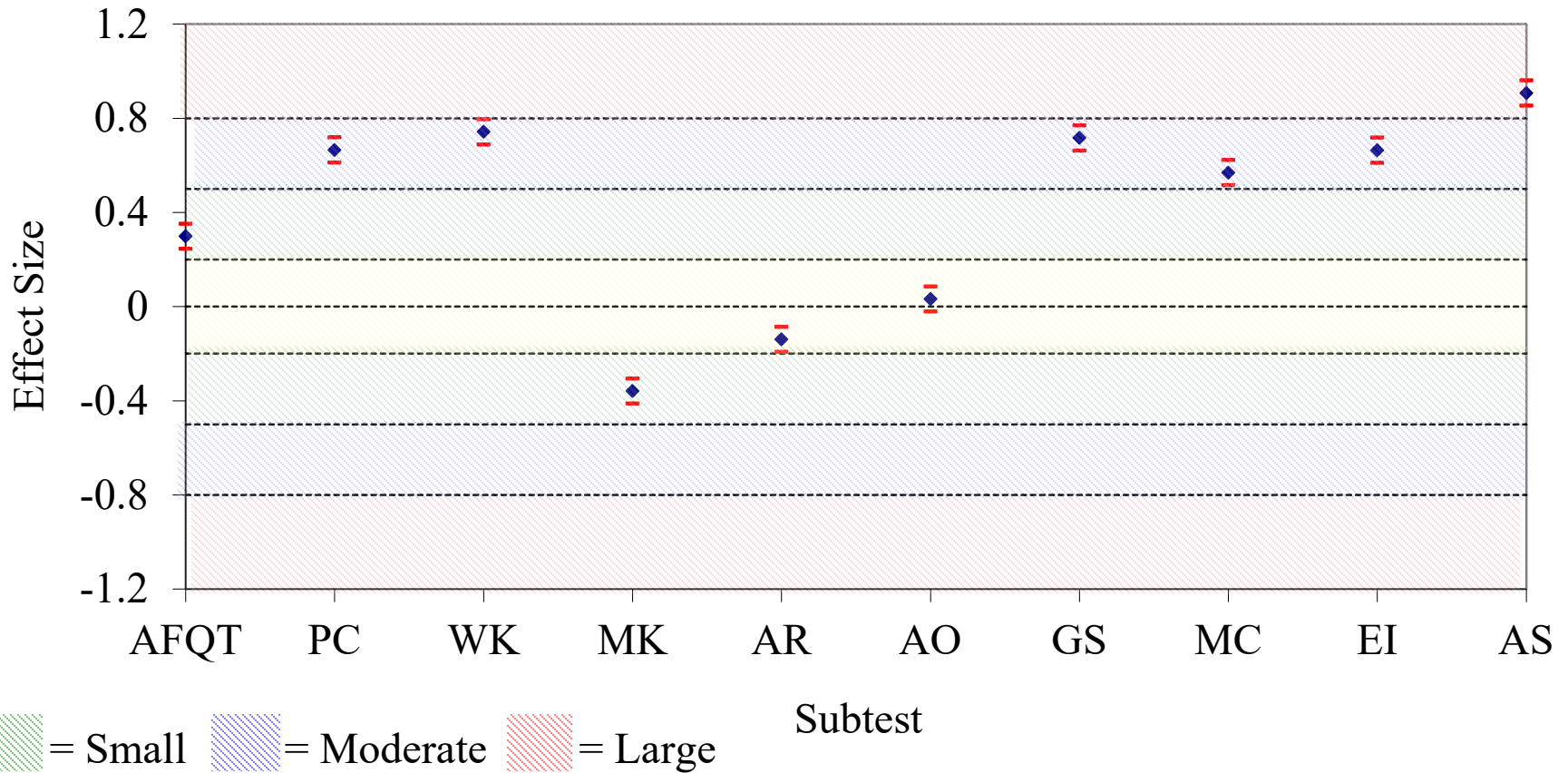
## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites Versus Hispanic Whites FY2021



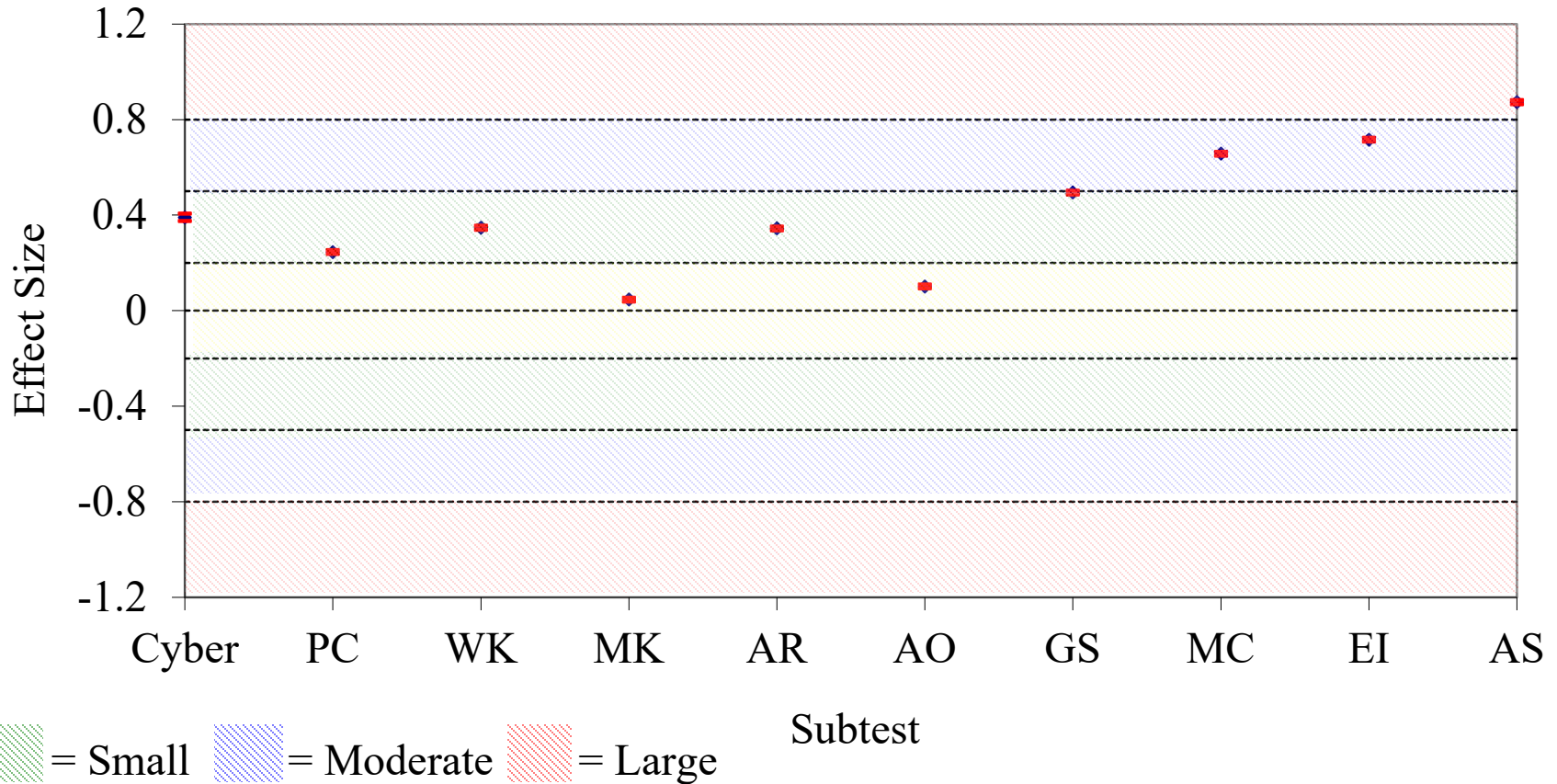
## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites Versus Non-Hispanic Blacks FY2021



## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites Versus Non-Hispanic Asians FY2021

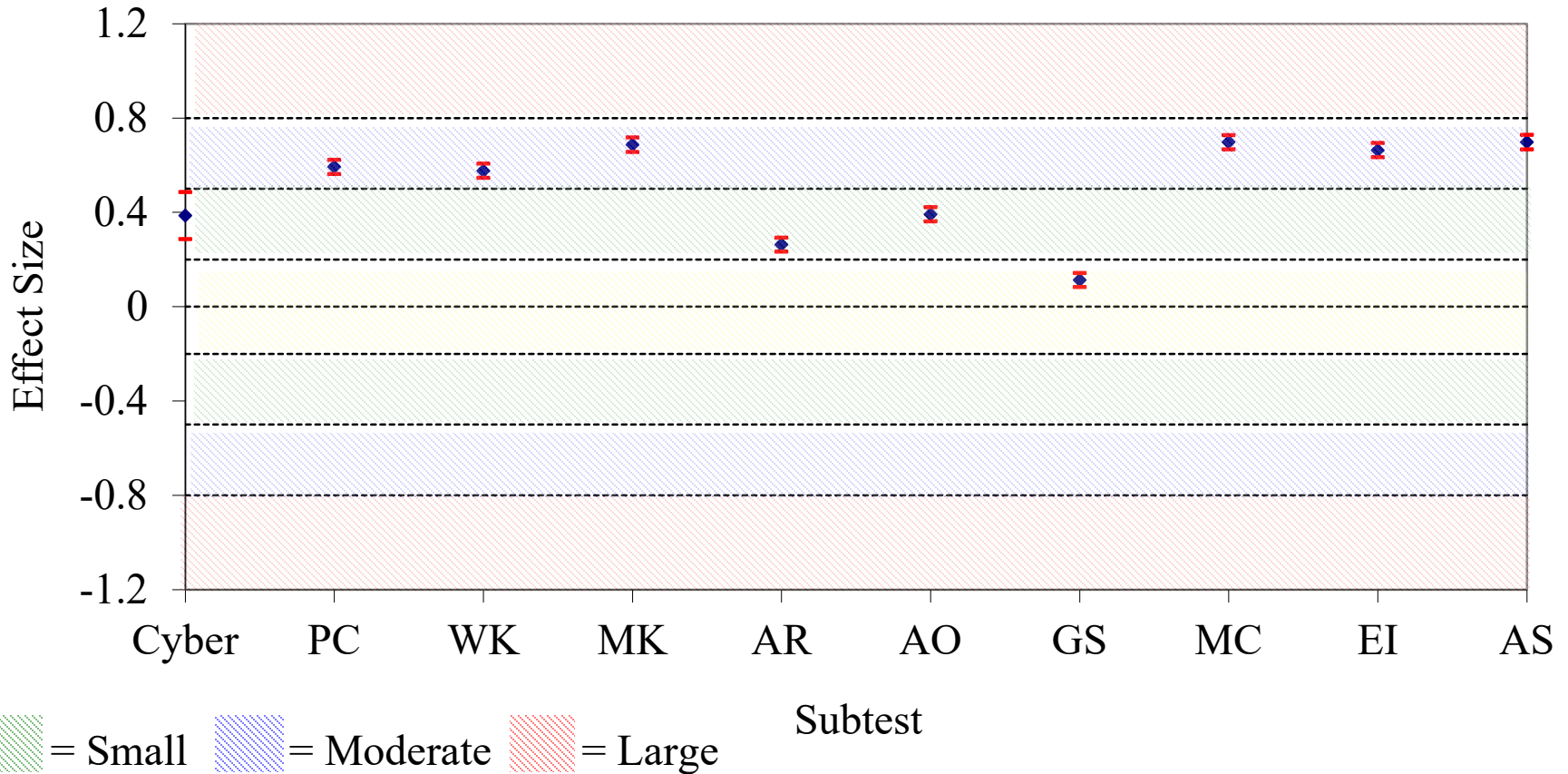


## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Males Versus Females – Cyber Test Sample FY2021

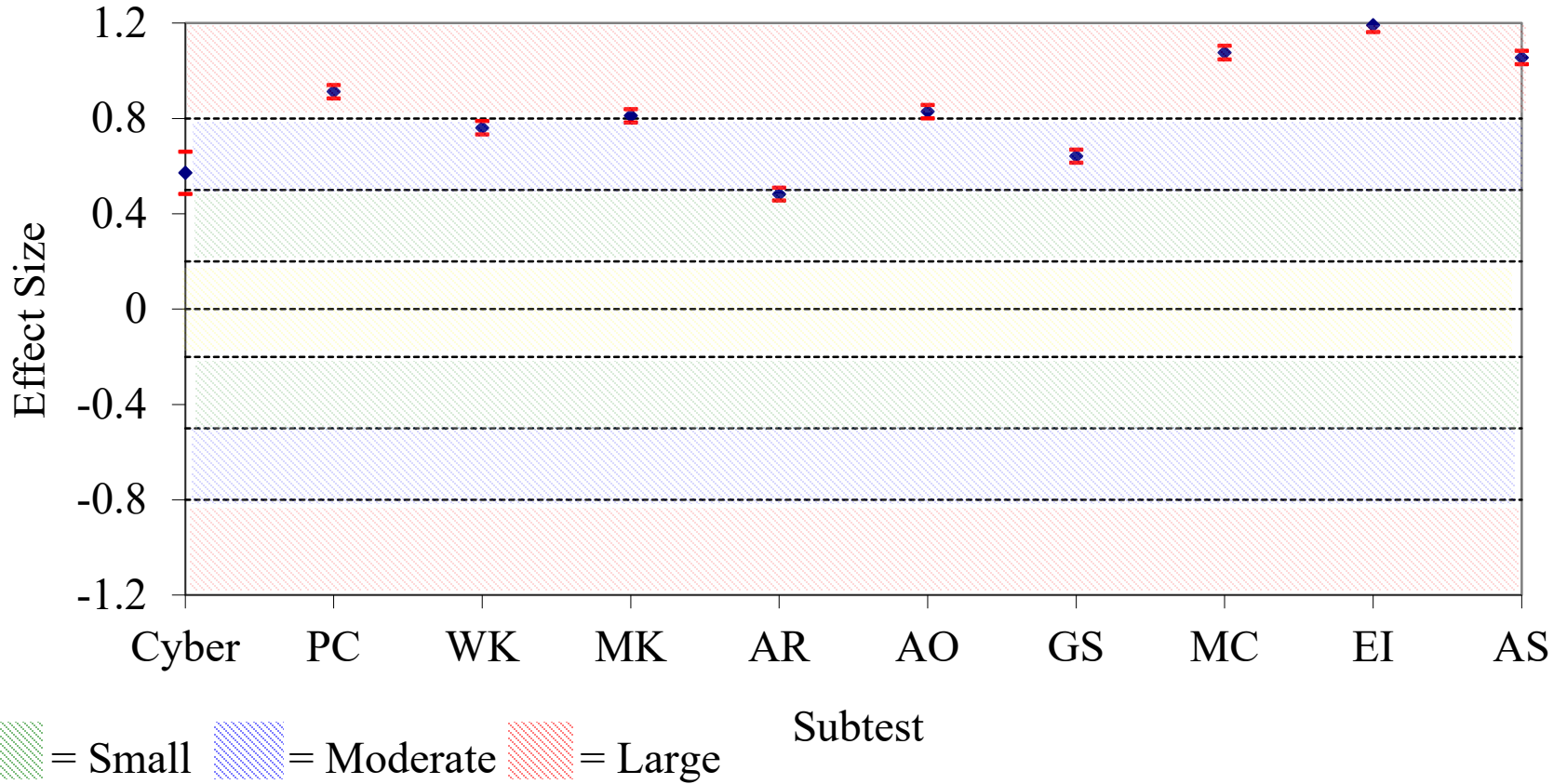




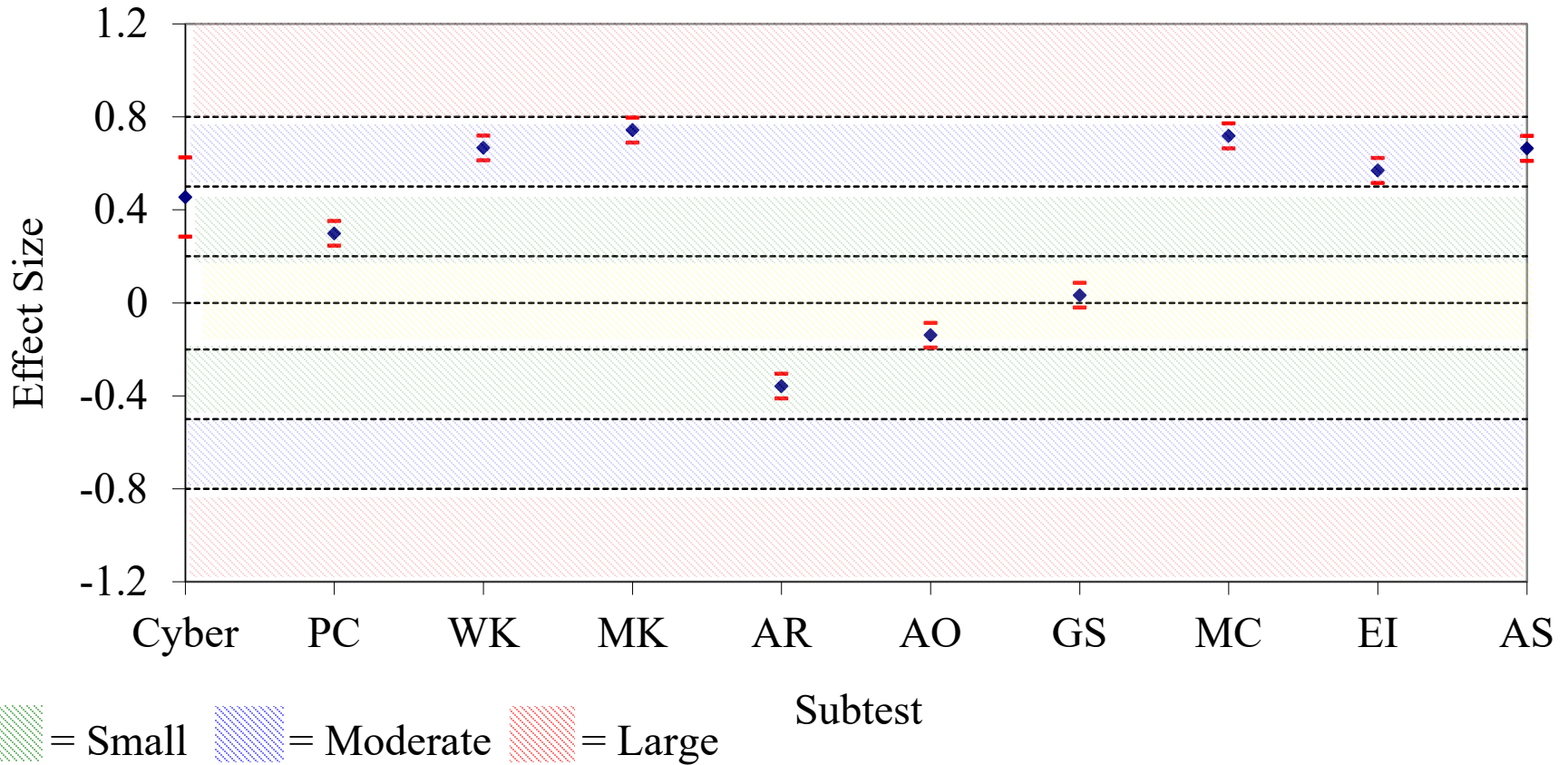
## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites Versus Hispanic Whites FY2021 – Cyber Test Sample



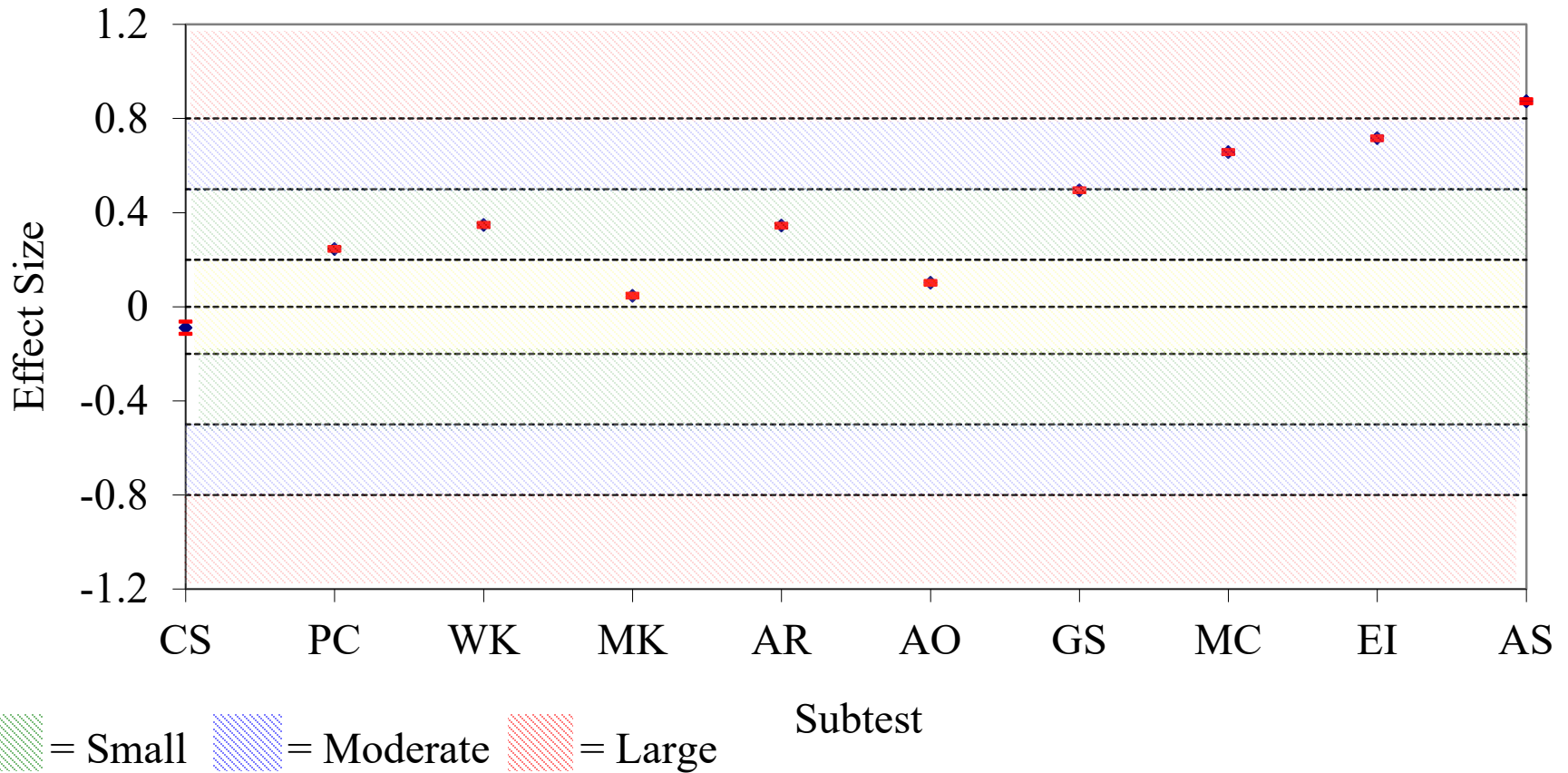
## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites Versus Non-Hispanic Blacks FY2021 – Cyber Test Sample



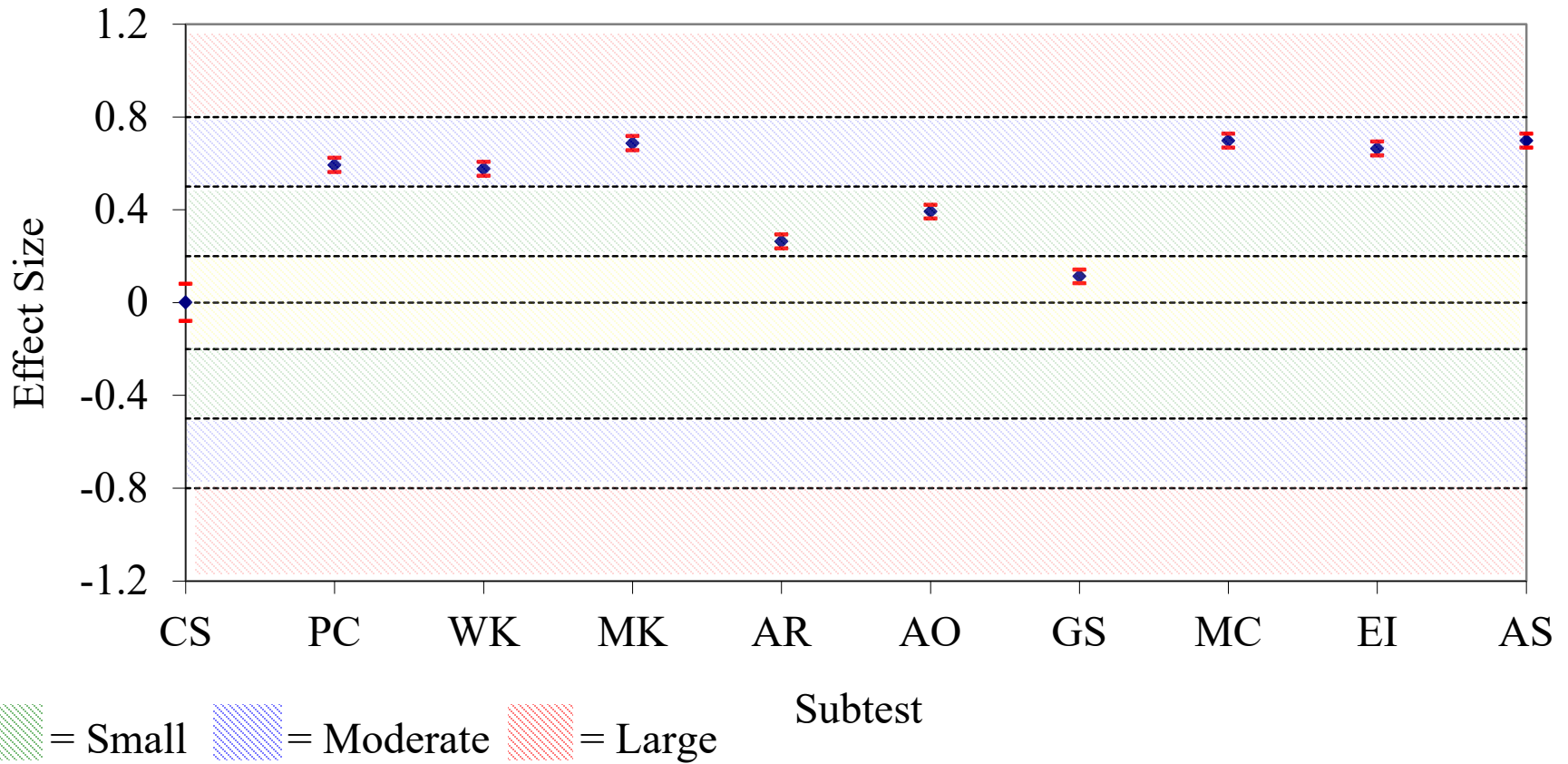
## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites Versus Non-Hispanic Asians FY2021 – Cyber Test Sample



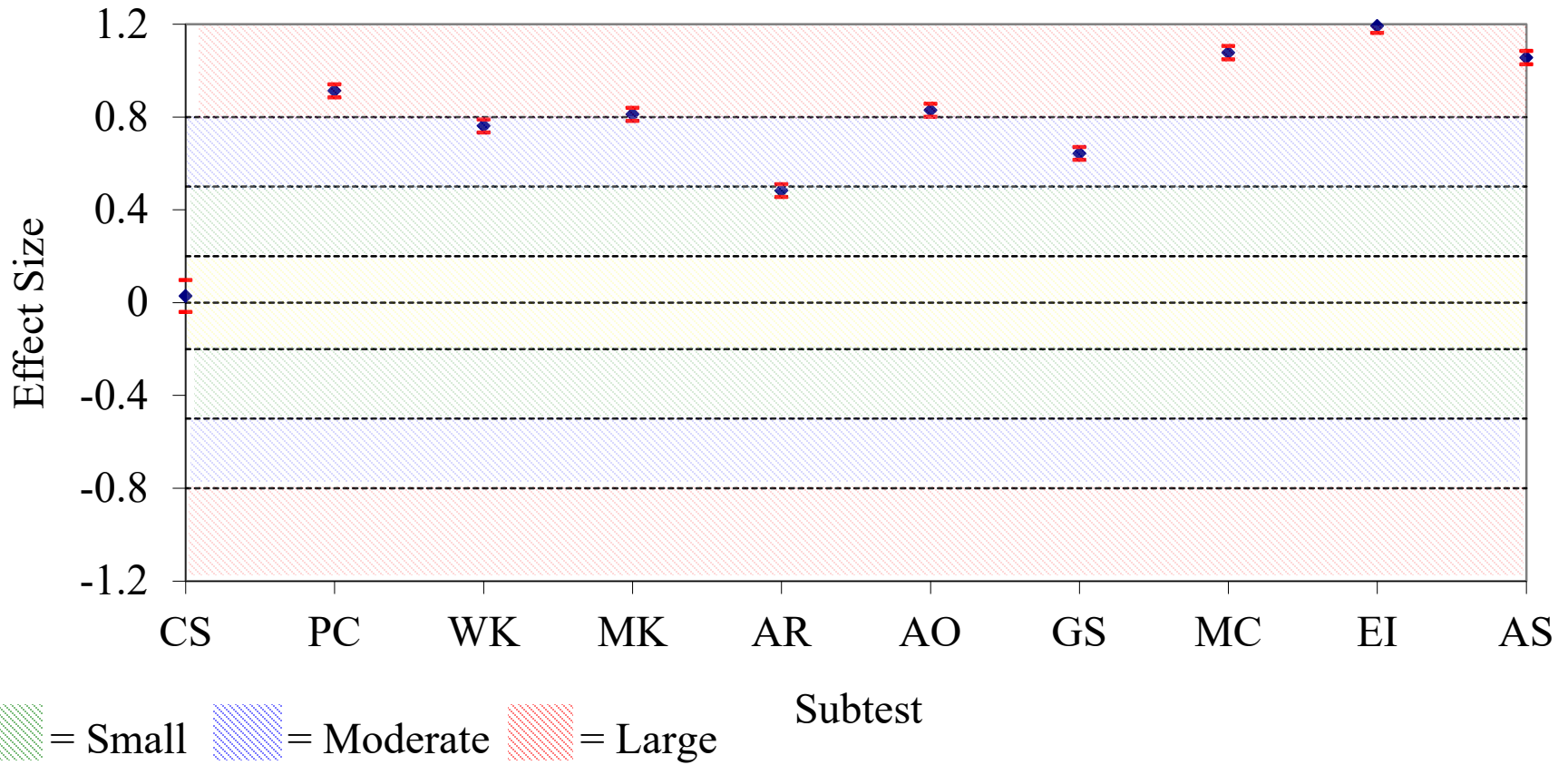
## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Males Versus Females – Coding Speed Sample FY2021



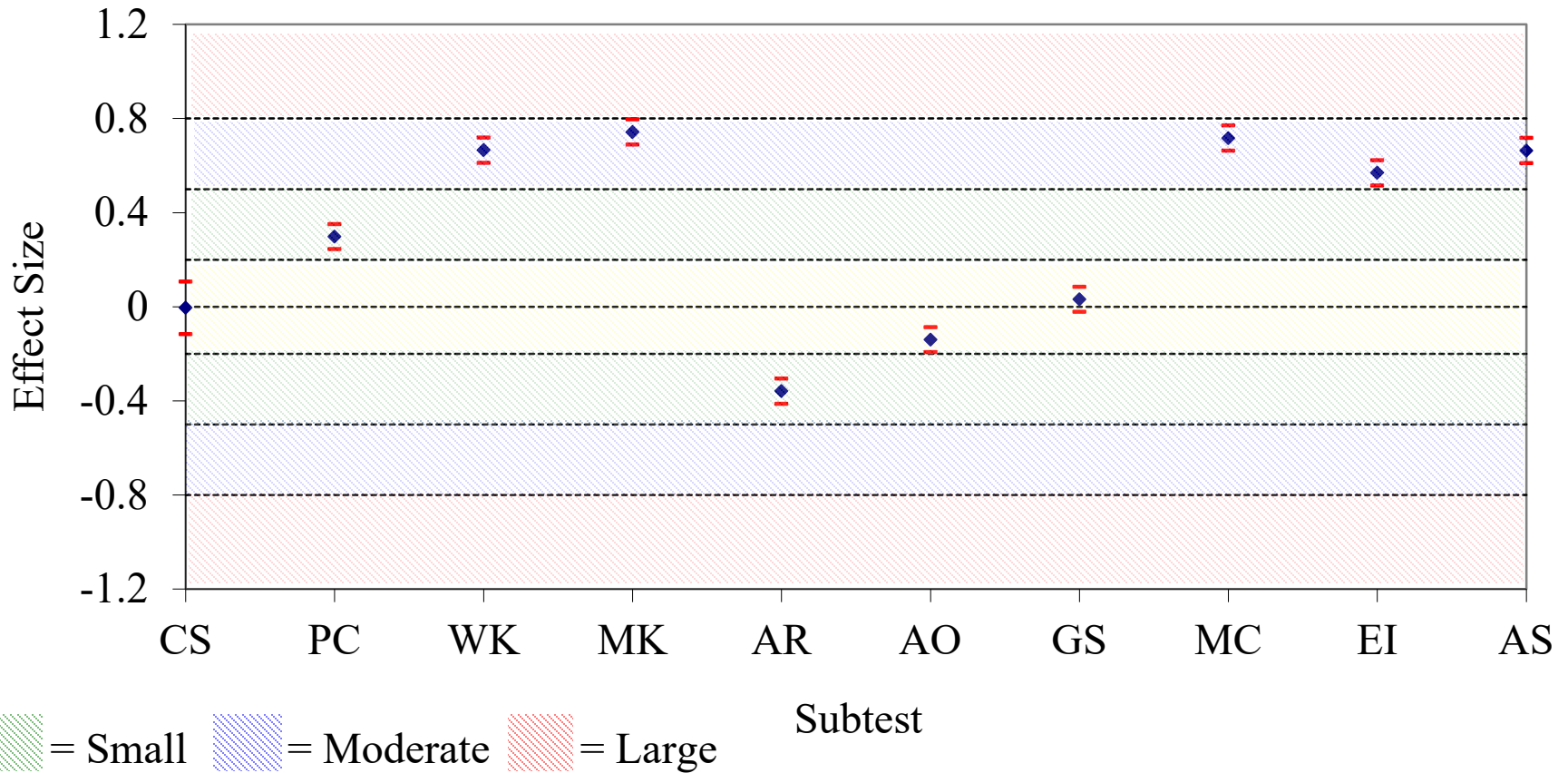
## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites Versus Hispanic Whites FY2021 – Coding Speed Sample



## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites Versus Non-Hispanic Blacks FY2021 – Coding Speed Sample



## Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites Versus Non-Hispanic Asians FY2021 – Coding Speed Sample



# CODING SPEED CONSIDERATIONS

- **Historically, CS has been very sensitive to changes in administration conditions, including:**
  - Shape of answer sheet bubbles (P&P administrations)
  - Type of input device (computer administrations)
  - [However, comparisons of rate scores across keyboard and mouse conditions in 2013 suggested no equating was needed at that time.]
- **If CS performance *is* impacted by administration conditions, use of non-standardized equipment (device, keyboard, mouse) to take iCAT could introduce construct irrelevant variance in CS scores.**
- **Potential fairness, accuracy, and validity concerns associated with CS:**
  - Could be affected by lag time in internet delivery (speeded test)
  - May not be feasible with touchscreen device
  - Could suffer from coachability, and susceptibility to invalid strategies that result in high scores