# ASVAB/AFQT Validity Framework

Briefing to the Defense Advisory Committee on Military Personnel Testing (DAC)
15 December 2022
Presented by Deirdre Knapp, HumRRO

66 Canal Center Plaza, Suite 700, Alexandria, VA 22314-1578 | Phone: 703.549.3611 | Fax: 703.549.9025 | www.humrro.org

# Briefing Agenda

- Purpose of the Validity Argument Frameworks
- Overview of Approach
- Overview of Associated Work
- AFQT and ASVAB Interpretive Arguments
- Report Format
- Summary of Work Products
- Concluding Observations

# Purpose of the Validity Argument Frameworks

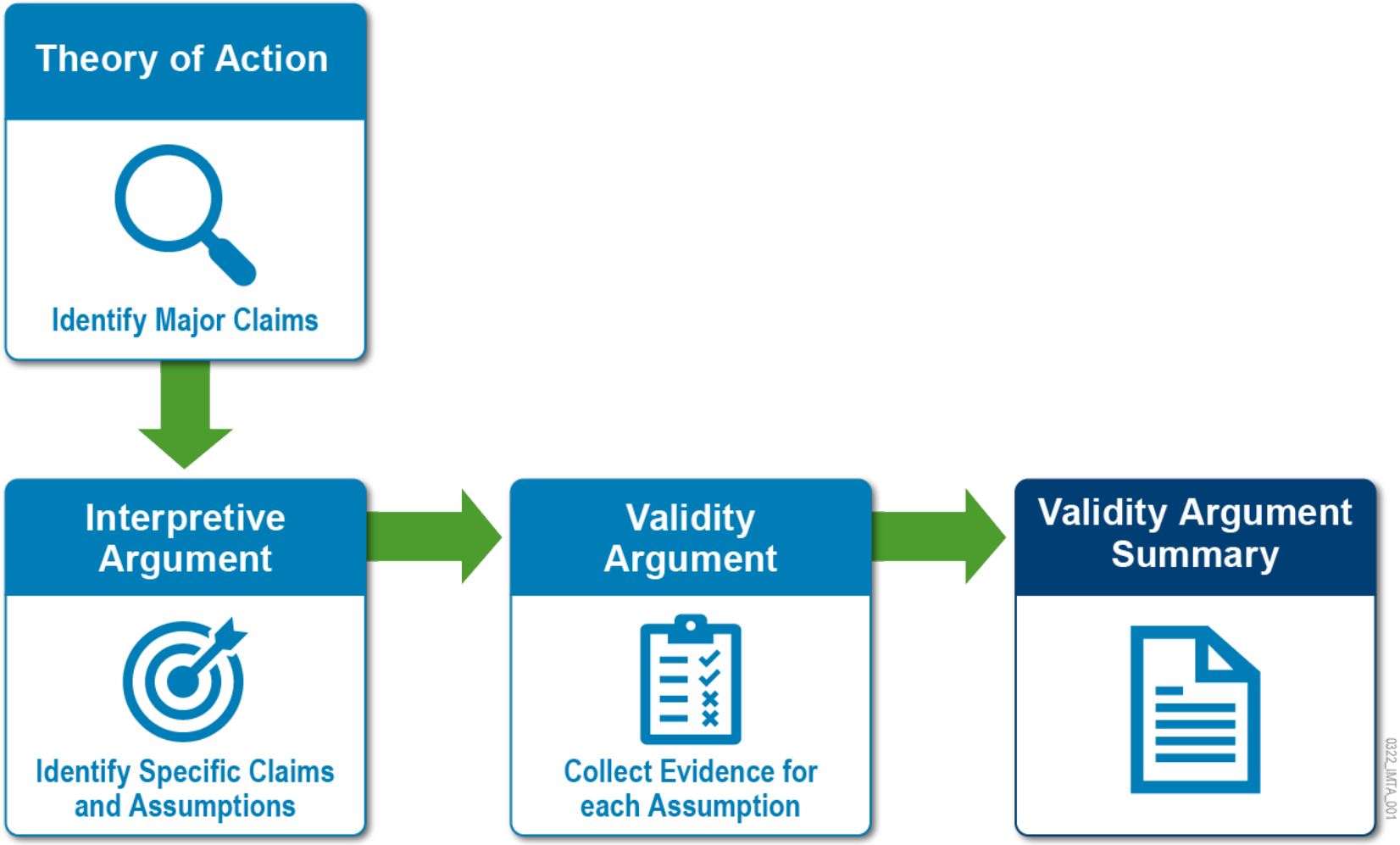
- Compile, organize, and review existing evidence related to the use of pre-enlistment assessments
  - Relevant information defined much more broadly than psychometric properties or criterion-related evidence
  - Includes all aspects of a measure's design, development, administration, score reporting, etc.
- Evaluate whether available evidence supports the use of these assessments for their intended purposes
- Identify ways to strengthen evidence supporting the use of these assessments
- Help inform improvements to these assessments in terms of content, scoring, administration, and/or interpretation

# Validity Argument Framework

***A theory of action is a useful starting point for developing a validity argument.***

- Theory of Action (TOA) — a connected set of propositions that explain a specific goal of the assessment. As such, TOAs are intimately linked to the purposes and uses of assessment scores.
- Interpretive Argument—a description of the inferences (i.e., claims and assumptions) that the assessment scores are intended to support.
- Validity Argument—evidence providing justification for the inferences in the interpretive argument.

# Validity Argument Method



# Overview of Associated Work

- First effort focused on use of ASVAB for enlisted selection (AFQT Validity Argument 1.0, May 2020)
- Second effort focused on use of TAPAS for enlisted selection and classification (citation)
- Validity argument for ASVAB use for classification decision-making (in progress)
- Revise/update AFQT validity argument 1.0 (in progress)

*Nature of this work is dynamic as validity argument evidence evolves over time.*

# AFQT Theory of Action

| AFQT Theory of Action – Selection | | |
|---|---|---|
| **Major Claims** | | |
| **I. *g* is broadly predictive of performance** → | **II. AFQT measures *g*** → | **III. Psychometric evidence supports the use of AFQT score categories for making selection decisions** |

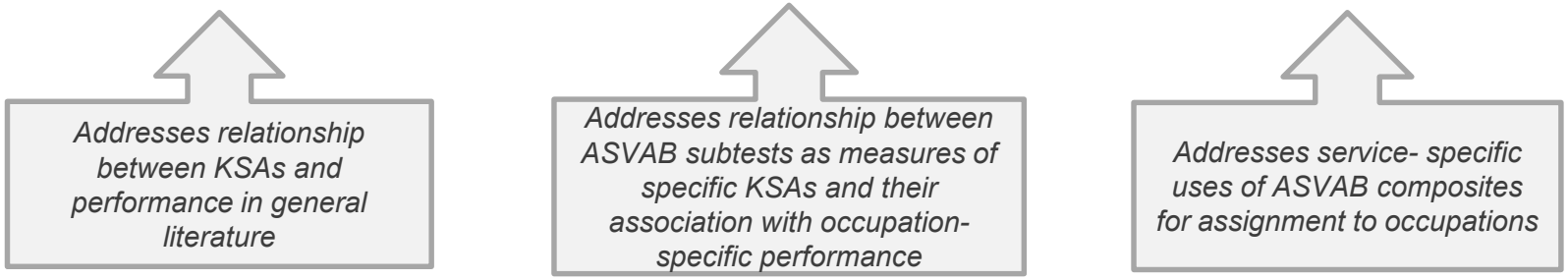
*Addresses relationship between g and performance in general literature*

*Addresses adequacy of AFQT as a measure of g*

*Addresses whether AFQT categories represent important differentiators among applicants*

# ASVAB Theory of Action

| ASVAB Theory of Action – Classification | | |
|---|---|---|
| **Major Claims** | | |
| **I. Specific KSAs are associated with occupation-specific performance and provide operational value for occupational classification.** → | **II. ASVAB subtests measure a useful sample of the KSAs associated with occupation-specific performance.** → | **III. Respondents classified on ASVAB composite scores have higher likelihood of success with particular military occupations.** ***(Service-specific)*** |



# AFQT Interpretive Argument Structure

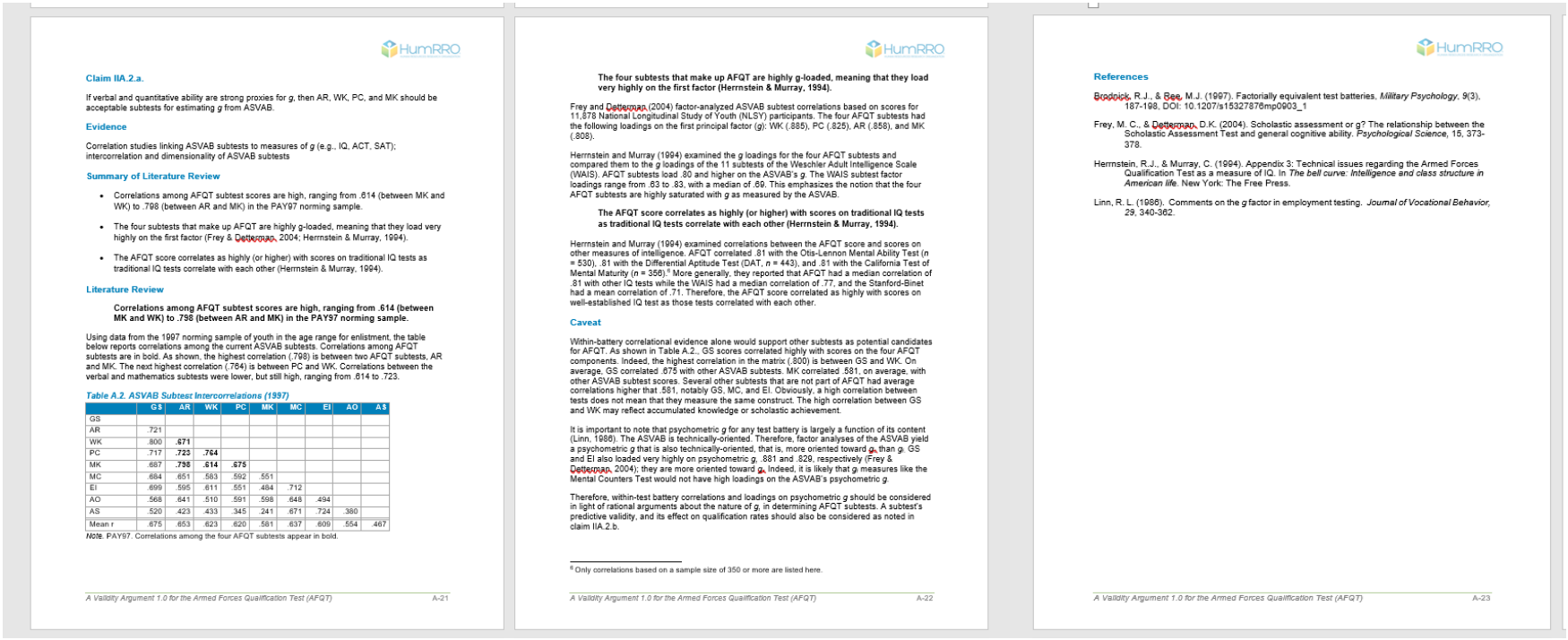
Example: Major Claim I: *g* is broadly predictive of performance

| Major Claim I |
| --- |
| I. *g* is broadly predictive of performance. |
| **Specific Claim** |
| I.1. *g* is a broad, stable construct that predicts performance. |
| **Assumptions** |
| I.1.a. If *g* is a broad, stable construct that predicts performance, then *g* can be measured by cognitive ability tests. |
| I.1.b. If *g* is a broad, stable construct that predicts performance, then *g* should be relatively stable over time. |
| I.1.c. If *g* is a broad, stable construct that predicts performance, then there should be a well-established body of validity evidence for *g* as a predictor of many performance outcomes (e.g., training/educational, job performance). |

# AFQT Interpretive Argument 2.0

- Revised to reflect improved approach/format adopted in ASVAB Classification interpretive argument
- Minimal impact on types of evidence to be collected
- Identified assumptions that are the same or largely the same between the AFQT and ASVAB interpretive arguments to simplify future updates to both validity arguments

# Evidence Summary Example from AFQT 1.0 Report



HumRRO

**Claim IIA.2.a.**

If verbal and quantitative ability are strong proxies for *g*, then AR, WK, PC, and MK should be acceptable subtests for estimating *g* from ASVAB.

**Evidence**

Correlation studies linking ASVAB subtests to measures of *g* (e.g., IQ, ACT, SAT); intercorrelation and dimensionality of ASVAB subtests

**Summary of Literature Review**

- Correlations among AFQT subtest scores are high, ranging from .614 (between MK and WK) to .798 (between AR and MK) in the PAY97 norming sample.
- The four subtests that make up AFQT are highly g-loaded, meaning that they load very highly on the first factor (Frey & Detterman, 2004; Herrnstein & Murray, 1994).
- The AFQT score correlates as highly (or higher) with scores on traditional IQ tests as traditional IQ tests correlate with each other (Herrnstein & Murray, 1994).

**Literature Review**

**Correlations among AFQT subtest scores are high, ranging from .614 (between MK and WK) to .798 (between AR and MK) in the PAY97 norming sample.**

Using data from the 1997 norming sample of youth in the age range for enlistment, the table below reports correlations among the current ASVAB subtests. Correlations among AFQT subtests are in bold. As shown, the highest correlation (.798) is between two AFQT subtests, AR and MK. The next highest correlation (.764) is between PC and WK. Correlations between the verbal and mathematics subtests were lower, but still high, ranging from .614 to .723.

*Table A.2. ASVAB Subtest Intercorrelations (1997)*

| | G[a] | AR | WK | PC | MK | MC | EI | AO | AS |
|---|---|---|---|---|---|---|---|---|---|
| GS | | | | | | | | | |
| AR | .721 | | | | | | | | |
| WK | .800 | **.671** | | | | | | | |
| PC | .717 | **.723** | **.764** | | | | | | |
| MK | .687 | **.798** | **.614** | **.675** | | | | | |
| MC | .684 | .651 | .583 | .592 | .551 | | | | |
| EI | .699 | .595 | .611 | .551 | .484 | .712 | | | |
| AO | .568 | .641 | .510 | .591 | .598 | .648 | .494 | | |
| AS | .520 | .423 | .433 | .345 | .241 | .671 | .724 | .380 | |
| Mean r | .675 | .653 | .623 | .620 | .581 | .637 | .609 | .554 | .467 |

*Note.* PAY97. Correlations among the four AFQT subtests appear in bold.



HumRRO

**The four subtests that make up AFQT are highly g-loaded, meaning that they load very highly on the first factor (Herrnstein & Murray, 1994).**

Frey and Detterman (2004) factor-analyzed ASVAB subtest correlations based on scores for 11,878 National Longitudinal Study of Youth (NLSY) participants. The four AFQT subtests had the following loadings on the first principal factor (*g*): WK (.885), PC (.825), AR (.858), and MK (.808).

Herrnstein and Murray (1994) examined the *g* loadings for the four AFQT subtests and compared them to the *g* loadings of the 11 subtests of the Weschler Adult Intelligence Scale (WAIS). AFQT subtests load .80 and higher on the ASVAB's *g*. The WAIS subtest factor loadings range from .63 to .83, with a median of .69. This emphasizes the notion that the four AFQT subtests are highly saturated with *g* as measured by the ASVAB.

**The AFQT score correlates as highly (or higher) with scores on traditional IQ tests as traditional IQ tests correlate with each other (Herrnstein & Murray, 1994).**

Herrnstein and Murray (1994) examined correlations between the AFQT score and scores on other measures of intelligence. AFQT correlated .81 with the Otis-Lennon Mental Ability Test (*n* = 530), .81 with the Differential Aptitude Test (DAT, *n* = 443), and .81 with the California Test of Mental Maturity (*n* = 356).[a] More generally, they reported that AFQT had a median correlation of .81 with other IQ tests while the WAIS had a median correlation of .77, and the Stanford-Binet had a mean correlation of .71. Therefore, the AFQT score correlated as highly with scores on well-established IQ test as those tests correlated with each other.

**Caveat**

Within-battery correlational evidence alone would support other subtests as potential candidates for AFQT. As shown in Table A.2., GS scores correlated highly with scores on the four AFQT components. Indeed, the highest correlation in the matrix (.800) is between GS and WK. On average, GS correlated .675 with other ASVAB subtests. MK correlated .581, on average, with other ASVAB subtest scores. Several other subtests that are not part of AFQT had average correlations higher that .581, notably GS, MC, and EI. Obviously, a high correlation between tests does not mean that they measure the same construct. The high correlation between GS and WK may reflect accumulated knowledge or scholastic achievement.

It is important to note that psychometric *g* for any test battery is largely a function of its content (Linn, 1988). The ASVAB is technically-oriented. Therefore, factor analyses of the ASVAB yield a psychometric *g* that is also technically-oriented, that is, more oriented toward *g*<sub>c</sub> than *g*<sub>f</sub>. GS and EI also loaded very highly on psychometric *g*, .881 and .829, respectively (Frey & Detterman, 2004); they are more oriented toward *g*<sub>c</sub>. Indeed, it is likely that *g*<sub>f</sub> measures like the Mental Counters Test would not have high loadings on the ASVAB's psychometric *g*.

Therefore, within-test battery correlations and loadings on psychometric *g* should be considered in light of rational arguments about the nature of *g*, in determining AFQT subtests. A subtest's predictive validity, and its effect on qualification rates should also be considered as noted in claim IIA.2.b.

[a] Only correlations based on a sample size of 350 or more are listed here.



HumRRO

**References**

Brodnick, R.J., & Bee, M.J. (1997). Factorially equivalent test batteries, *Military Psychology, 9*(3), 187-198, DOI: 10.1207/s15327876mp0903_1

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, 373-378.

Herrnstein, R.J., & Murray, C. (1994). Appendix 3: Technical issues regarding the Armed Forces Qualification Test as a measure of IQ. In *The bell curve: Intelligence and class structure in American life*. New York: The Free Press.

Linn, R. L. (1988). Comments on the *g* factor in employment testing. *Journal of Vocational Behavior, 29*, 340-362.

# Report Format

- Description of the assessment (e.g., AFQT) and the validity argument approach
- Table for each major claim the lists the associated specific claims and assumptions; each assumption is hyperlinked to the applicable evidence summary
- Narrative summary of the evidence for each major claim
- Recommendations
  - For the AFQT v1.0 report, recommendations made in three areas
    - Philosophy underlying test content decisions
    - Suggestions for further research
    - Administrative improvements
  - Ongoing progress being made on several of these recommendations

# Summary of Work Products

- Recommendations to help inform future R&D (including next generation ASVAB) and operational practices, as applicable
- Organized archive of nearly 500 references (e.g., technical reports, briefings, journal pubs, books, notes from interviews with DTAC and service representatives)
- Dozens of succinct research summaries on various topics to serve as resources to DTAC and service researchers
- Organizing framework into which new evidence can be placed and communicated
- Detailed process map and supporting tools for periodically updating each validity argument

# Concluding Observations

- Challenging to develop clear, cohesive, and comprehensive interpretive arguments
- Need to consider primary audience for the evidence summaries
- Evidence summaries for individual assumptions benefited from many layers of review
- We did not try to quantify evaluation of the evidence and would be hesitant to do so

# Validity Argument Project Leads/Advisors

- Laura Ford
- Andrea Sinclair
- Art Thacker
- Deirdre Knapp
- Teresa Russell