Evaluating AFQT Scores for Differential Prediction

Presentation to the Defense Advisory Committee on Military Personnel Testing (DACMPT)

December 16, 2022

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

Agenda

- Overview
- Historical findings
- The current study
  - Sample and data
  - Analyses
  - Service-specific results
  - Attrition results
- Summary, future research, and questions for the DAC

Overview

- Professional standards for psychological measurement recommend test scores used for selection decisions (such as the AFQT) be evaluated for *differential prediction.*

- ***What is differential prediction?***

  "The systematic under- or over-prediction of criterion performance for people belonging to *subgroups* differentiated by characteristics not relevant to criterion performance" (Society for Industrial and Organizational Psychology, 2018).

- When evaluating differential prediction in the context of personnel selection, these *subgroups* are often defined by test-takers' race/ethnicity or gender (*biological sex*).

Interpretation of over- and under-prediction in the literature

- Historically, the literature on differential prediction has been primarily concerned with the under-prediction of performance for the focal subgroup (e.g., Black applicants, female applicants)—with respect to **positively valenced outcomes** such as job performance and training success.
  - If a common regression line results in **under-prediction** of performance relative to a focal subgroup's regression line, *use of the common regression line disadvantages that subgroup*.
  - If a common regression line results in **over-prediction** of performance relative to a focal subgroup's regression line, *use of the common regression line does not disadvantage that subgroup*.

- Note that these interpretations are flipped for **negatively valenced outcomes** (such as attrition).
  - If a common regression line results in **under-prediction** of attrition relative to a focal subgroup's regression line, *use of the common regression line does not disadvantage that subgroup*.
  - If a common regression line results in **over-prediction** of attrition relative to a focal subgroup's regression line, *use of the common regression line disadvantages that subgroup*.

The current study

- Using five years of data on enlisted applicants who completed the ASVAB as part of the Enlisted Testing Program (ETP), we evaluated whether AFQT scores exhibited differential prediction for predicting multiple training performance and retention criteria, within multiple Services, for hundreds of enlisted military jobs/training courses.

Historical Findings

Past research

- Race/Ethnicity
  - Much past research in civilian and military settings has examined cognitive ability test scores for differential prediction with respect to race/ethnicity – with most focused on White-Black comparisons.
  - General findings have been that use of a common regression line for cognitive ability tests will tend to over-predict performance for Black individuals and under-predict performance for White individuals relative to use of subgroup-specific regression lines (civilian research– e.g., Berry, 2015, ASVAB research – e.g., Wise et al., 1992).

- Gender (*Biological Sex)*
  - Compared to past research on cognitive ability test scores for differential prediction with respect to race/ethnicity, research with regard to gender has been less consistent.
  - Past research has indicated cognitive ability tests sometimes under-predict females' performance, particularly when college grades are the criterion of interest, but in the Wise et al. (1992) study, where ASVAB technical subtests and technical performance were of primary interest, female performance was either over-predicted or predicted at a comparable level to males.

Important caveats about past research

- Despite the amount of research to date, much of the civilian employment research in this area is based on older data (1980s and earlier) and predates the recognition of the methodological concerns raised over the past decade pertaining to the traditional implementation of the Cleary approach (Aguinis, Culpepper, & Pierce, 2010, Berry, 2015, Berry & Zhao, 2015; Mattern & Patterson, 2013).
  - See the "primer" at the end this presentation for further details on the Cleary approach (*Back Up Slides Part 1: Primer on Evaluating Test Scores for Differential Prediction*).

- In light of critiques of past differential prediction research by Aguinis, Berry, and others in the field, for this study, we adopted an expanded version of the Cleary approach that accounts for the presence of range restriction stemming from the use of ASVAB for general enlistment and occupation qualification decisions.
  - See the "Methods-Related Details for AFQT Differential Prediction Analyses" read-ahead document.

The Current Study

Sample and data

- We analyzed AFQT scores and demographic data for 1,603,749 individuals who completed the ASVAB as part of the Enlisted Testing Program (ETP) between 1 October 2013 and 30 September 2018 (i.e., FY14-FY18 applicants).
- For the subset of these applicants who accessed into Regular Air Force, Army, Coast Guard, Marine Corps, and Navy, we obtained accession and separation data current through November 2021 allowing us to calculate a 36-month attrition criterion.
- We also obtained additional criterion data for subsets of Air Force, Army, Marine Corps, and Navy accessions from archival data sources from each of those Services:
  - **Air Force**: Awarding course grades for 60K+ Airmen as they completed awarding courses in over 60+ Air Force Specialties (AFS).
  - **Army**: Army-wide job knowledge test (JKT) scores for 38K+ Soldiers from 30+ Military Occupational Specialties (MOS), and MOS-specific JKT scores from 27K+ Soldiers from 10 MOS as they exited initial military training (IMT).
  - **Marine Corps:** Initial military training course graduation indicators (i.e., graduated course without a setback) for 158K+ Marines across 120+ training courses.
  - **Navy**: Initial technical training graduation indicators (i.e., graduated without a setback) for 68K+ Sailors across 40+ Ratings.

Analyses

- We formally evaluated whether AFQT scores exhibited differential prediction, and if so, determined what type of differences they exhibited (i.e., intercept or slope differences) using an updated version of the Cleary-based approach described earlier that accounts for selection-related artifacts raised as concerns in past research. Analyses were conducted for four subgroup contrasts:
  - White non-Hispanic vs. Black non-Hispanic
  - White non-Hispanic vs. Hispanic White
  - White non-Hispanic vs. Asian non-Hispanic
  - Male vs. Female
- We estimated the magnitude of differences in prediction for subgroup-specific AFQT regression models by calculating d_{Mod} statistics (Nye & Sackett, 2016; Dahlke & Sackett, 2018).
- We performed post hoc analyses to estimate the power we had to detect statistically significant differences between the models being compared under the updated Cleary-based approach.

BLUF: Bottom line up front

Service-Specific Training and Job Knowledge Criteria

- Across Services, we examined 664 military occupation/training course-by-subgroup contrast combinations for evidence of differential prediction for the AFQT for predicting Service-specific criteria. Most combinations examined (79.1%) yielded no statistically significant evidence of differential prediction.
- Consistent with past research in the civilian and military cognitive ability testing literature, use of a common regression line for the AFQT tended to over-predict performance for Black individuals and under-predict performance for White individuals relative to use of subgroup-specific regression lines (e.g., Berry, 2015, Wise et al., 1992). Findings for other focal subgroups reflected a mix of over- and under-prediction when differences were found.
- Across Services, of the 139 military occupation/training course-by-subgroup contrast combinations that exhibited statistically significant evidence of differential prediction for the AFQT, most exhibited **small or very small** prediction differences at AFQT Category IIIB lower bound (31) and AFQT Category IIIA lower bound (50).

Attrition Criterion

- Relative to the Service-specific criteria, there was more evidence of differential prediction of the AFQT for predicting probabilities of 36-month attrition (caveat – AFQT was not designed to predict attrition).
- Prediction differences for attrition were particularly strong for Hispanic White and Asian non-Hispanic Servicemembers relative to White non-Hispanic Servicemembers, with evidence that a common regression line would result in relatively large over-prediction of attrition for Hispanic White and Asian non-Hispanic Servicemembers (particularly in the Army).
  - Differences suggest another reason why it is important to consider non-cognitive measures (e.g., TAPAS) if one's objective is to predict first term attrition with a selection/classification measure.

Differential Prediction Results
Air Force

Differential prediction outcome summary – AFS awarding course grades

	WNH-BNH		WNH-HW		WNH-ANH		M-F		Grand Totals	
# of AFS	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Total	54		51		10		50		165	
No Difference	37	68.5	45	88.2	6	60.0	38	76.0	126	76.4
Slope Difference										
Over	1	1.9	0	0.0	0	0.0	0	0.0	1	0.6
Under	6	11.1	2	3.9	2	20.0	3	6.0	13	7.9
Intercept Difference										
Over	10	18.5	3	5.9	0	0.0	1	2.0	14	8.5
Under	0	0.0	1	2.0	2	20.0	8	16.0	11	6.7

Over	Indicates over-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).
Under	Indicates under-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).

- AFQT exhibited no statistically significant evidence of differential prediction for the vast majority of AFS examined with respect to predicting awarding course grades.
- Consistent with past research, when intercept differences were found, they suggested use of a common regression line would result in over-prediction of course grades for Black Airmen.
- Also of note, when prediction differences were found for Males vs. Females, results largely suggested use of a common regression line would result in under-prediction of course grades for Females.

Summarizing the magnitude of prediction differences

Magnitude / Direction of Effect
Large overprediction ($d_{Mod} \geq .80$)
Moderate overprediction ($.50 \leq d_{Mod} < .80$)
Small overprediction ($.20 \leq d_{Mod} < .50$)
Very small overprediction ($0 \leq d_{Mod} < .20$)
Very small underprediction ($-.20 < d_{Mod} < 0$)
Small underprediction ($-.50 < d_{Mod} \leq -.20$)
Moderate underprediction ($-.80 < d_{Mod} \leq -.50$)
Large underprediction ($d_{Mod} \leq -.80$)

- d_{Mod} provides an index of the standardized difference between predictions from subgroup-specific regression lines at a given point on the AFQT score distribution.
- For purposes of the summaries that follow, we've classified d_{Mod} using Cohen's conventions for judging the magnitude of standardized mean differences using Cohen's *d* statistics:
  - Large ($|d| \geq .80$), moderate ($|d| \geq .50$), small ($|d| \geq .20$).
- **Caution:** What we've provided is a simple categorization of d_{Mod} effect sizes by magnitude. Ultimately, what constitutes a large/moderate/small effect is context-dependent and up to DoD and its stakeholders to judge.
- We focus on differences in prediction at two key points on the AFQT score distribution:
  - Lower bound of Cat IIIB (31) – general enlistment qualification
  - Lower bound of Cat IIIA (50) – general incentives qualification

Magnitude of prediction differences: d_{Mod} summary – AFS awarding course grades

	At AFQT Category IIIB Lower Bound (31)				At AFQT Category IIIA Lower Bound (50)			
Magnitude / Direction of Effect	WNH-BNH	WNH-HW	WNH-ANH	M-F	WNH-BNH	WNH-HW	WNH-ANH	M-F
Large overprediction ($d_{Mod} \geq .80$)	2							
Moderate overprediction ($.50 \leq d_{Mod} < .80$)					3			
Small overprediction ($.20 \leq d_{Mod} < .50$)	3			1	8	1		
Very small overprediction ($0 \leq d_{Mod} < .20$)	6	3			2	2		1
Very small underprediction ($-.20 < d_{Mod} < 0$)	2	1		1	1	2	1	3
Small underprediction ($-.50 < d_{Mod} \leq -.20$)			1	8	2		1	7
Moderate underprediction ($-.80 < d_{Mod} \leq -.50$)	3	1	2	1			2	
Large underprediction ($d_{Mod} \leq -.80$)	1	1	1	1	1	1		1
Totals	17	6	4	12	17	6	4	12

- Table shows the magnitude of differences in prediction for AFS that showed statistically significant evidence of differential prediction for AFQT scores.
- Cells show number of AFS that exhibited differences in prediction of the given d_{Mod} size at AFQT Category IIIB lower bound (31) or AFQT Category IIIA (50).
- Most prediction differences were **small or very small** at both of these AFQT score levels.

Differential Prediction Results
Army

Differential prediction outcome summary – Army-wide JKT

	WNH-BNH		WNH-HW		WNH-ANH		M-F		Grand Totals	
# of MOS	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Total	22		28		11		20		81	
No Difference	5	22.7	21	75.0	8	72.7	17	85.0	51	63.0
Slope Difference										
Over	1	4.5	1	3.6	0	0.0	0	0.0	2	2.5
Under	1	4.5	1	3.6	1	9.1	0	0.0	3	3.7
Intercept Difference										
Over	13	59.1	4	14.3	2	18.2	3	15.0	22	27.2
Under	2	9.1	1	3.6	0	0.0	0	0.0	3	3.7

Over	Indicates over-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).
Under	Indicates under-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).

- AFQT exhibited statistically significant evidence of differential prediction for the majority of MOS examined with respect to predicting Army-wide job knowledge test (JKT) scores for the WNH vs. BNH Soldier contrast.
- Consistent with past research, when intercept differences were found, they suggested use of a common regression line would result in over-prediction of Army-wide JKT scores for Black Soldiers.
- AFQT exhibited no statistically significant evidence of differential prediction for the vast majority of MOS examined with respect to predicting Army-wide JKT scores for all other subgroup contrasts examined, but when differences were found, they tended to indicate over-prediction of JKT scores for the focal groups of interest (i.e.,. Hispanic White, Asian non-Hispanic, Female).

Magnitude of prediction differences: d_{Mod} summary – Army-wide JKT

	At AFQT Category IIIB Lower Bound (31)				At AFQT Category IIIA Lower Bound (50)			
Magnitude / Direction of Effect	WNH-BNH	WNH-HW	WNH-ANH	M-F	WNH-BNH	WNH-HW	WNH-ANH	M-F
Large overprediction ($d_{Mod} \geq .80$)	1							
Moderate overprediction ($.50 \leq d_{Mod} < .80$)	3	2			3	1		
Small overprediction ($.20 \leq d_{Mod} < .50$)	4		1	1	10	2	2	
Very small overprediction ($0 \leq d_{Mod} < .20$)	6	3	1	2	4	3	1	3
Very small underprediction ($-.20 < d_{Mod} < 0$)	2	1	1					
Small underprediction ($-.50 < d_{Mod} \leq -.20$)	1							
Moderate underprediction ($-.80 < d_{Mod} \leq -.50$)						1		
Large underprediction ($d_{Mod} \leq -.80$)		1						
Totals	17	7	3	3	17	7	3	3

- Table shows the magnitude of differences in prediction for MOS that showed statistically significant evidence of differential prediction for AFQT scores.
- Cells show number of MOS that exhibited differences in prediction of the given d_{Mod} size at AFQT Category IIIB lower bound (31) or AFQT Category IIIA (50).
- Like the Air Force results, most prediction differences were **small or very small** at both of these AFQT score levels, with a trend toward over-prediction for the focal group.

Differential prediction outcome summary – MOS-specific JKT

	WNH-BNH		WNH-HW		WNH-ANH		M-F		Grand Totals	
# of MOS	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Total	10		10		7		6		33	
No Difference	2	20.0	4	40.0	4	57.1	2	33.3	12	36.4
Slope Difference										
Over	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Under	0	0.0	0	0.0	1	14.3	0	0.0	1	3.0
Intercept Difference										
Over	8	80.0	5	50.0	2	28.6	3	50.0	18	54.5
Under	0	0.0	1	10.0	0	0.0	1	16.7	2	6.1

Over	Indicates over-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).
Under	Indicates under-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).

- AFQT exhibited statistically significant evidence of differential prediction for the majority of MOS examined with respect to predicting MOS-specific JKT scores for all subgroup contrasts examined except WNH vs. ANH. The largest proportion of differences were found for the WNH vs. BNH Soldier contrast.
- Consistent with past research, when intercept differences were found, they suggested use of a common regression line would result in over-prediction of MOS-specific JKT scores for Black Soldiers.
- When differences were found for other subgroup contrasts, they also tended to indicate over-prediction of JKT scores for the focal groups of interest (i.e., Hispanic White, Asian non-Hispanic, Female).

Magnitude of prediction differences: d_{Mod} summary – MOS-specific JKT

	At AFQT Category IIIB Lower Bound (31)				At AFQT Category IIIA Lower Bound (50)			
Magnitude / Direction of Effect	WNH-BNH	WNH-HW	WNH-ANH	M-F	WNH-BNH	WNH-HW	WNH-ANH	M-F
Large overprediction ($d_{Mod} \geq .80$)								
Moderate overprediction ($.50 \leq d_{Mod} < .80$)	2				1			
Small overprediction ($.20 \leq d_{Mod} < .50$)	6	3	2	3	7	4	2	3
Very small overprediction ($0 \leq d_{Mod} < .20$)		2				1	1	
Very small underprediction ($-.20 < d_{Mod} < 0$)			1			1		1
Small underprediction ($-.50 < d_{Mod} \leq -.20$)		1		1				
Moderate underprediction ($-.80 < d_{Mod} \leq -.50$)								
Large underprediction ($d_{Mod} \leq -.80$)								
Totals	8	6	3	4	8	6	3	4

- Table shows the magnitude of differences in prediction for MOS that showed statistically significant evidence of differential prediction for AFQT scores.
- Cells show number of MOS that exhibited differences in prediction of the given d_{Mod} size at AFQT Category IIIB lower bound (31) or AFQT Category IIIA (50).
- Like the Army-wide JKT results, most prediction differences were **small or very small** at both of these AFQT score levels, with a trend toward over-prediction for the focal group.

Differential Prediction Results
Marine Corps

Differential prediction outcome summary – probability of training course graduation

	WNH-BNH		WNH-HW		WNH-ANH		M-F		Grand Totals	
# of Courses	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Total	63		119		26		66		274	
No Difference	50	79.4	105	88.2	22	84.6	55	83.3	232	84.7
Slope Difference										
Over	1	1.6	1	0.8	0	0.0	3	4.5	5	1.8
Under	1	1.6	6	5.0	2	7.7	2	3.0	11	4.0
Intercept Difference										
Over	9	14.3	4	3.4	1	3.8	3	4.5	17	6.2
Under	2	3.2	3	2.5	1	3.8	3	4.5	9	3.3

Over	Indicates over-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).
Under	Indicates under-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).

- AFQT exhibited no statistically significant evidence of differential prediction for the vast majority of training courses examined with respect to graduation probability.
- Consistent with past research, when intercept differences were found, they suggested use of a common regression line would result in over-prediction of training graduation probabilities for Black Marines.

Magnitude of prediction differences: d_{Mod} summary – probability of training course graduation

	At AFQT Category IIIB Lower Bound (31)				At AFQT Category IIIA Lower Bound (50)			
Magnitude / Direction of Effect	WNH-BNH	WNH-HW	WNH-ANH	M-F	WNH-BNH	WNH-HW	WNH-ANH	M-F
Large overprediction ($d_{Mod} \geq .80$)	2	2		3	2	2		3
Moderate overprediction ($.50 \leq d_{Mod} < .80$)	1	1			1			
Small overprediction ($.20 \leq d_{Mod} < .50$)	4	1	1	2	6		1	2
Very small overprediction ($0 \leq d_{Mod} < .20$)	3	1		1	4	4		1
Very small underprediction ($-.20 < d_{Mod} < 0$)	2	3	1			6	2	1
Small underprediction ($-.50 < d_{Mod} \leq -.20$)	1	3	1	1		2		3
Moderate underprediction ($-.80 < d_{Mod} \leq -.50$)		2	1	3			1	1
Large underprediction ($d_{Mod} \leq -.80$)		1		1				
Totals	13	14	4	11	13	14	4	11

- Table shows the magnitude of differences in prediction for training courses that showed statistically significant evidence of differential prediction for AFQT scores.
- Cells show number of training courses that exhibited differences in prediction of the given d_{Mod} size at AFQT Category IIIB lower bound (31) or AFQT Category IIIA (50).
- Like results for the Air Force and Army, most prediction differences were **small or very small** at both of these AFQT score levels.

Differential Prediction Results
Navy

Differential prediction outcome summary – probability of training graduation

	WNH-BNH		WNH-HW		WNH-ANH		M-F		Grand Totals	
# of Ratings	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Total	29		28		16		38		111	
No Difference	27	93.1	28	100.0	14	87.5	35	92.1	104	93.7
Slope Difference										
Over	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Under	1	3.4	0	0.0	1	6.3	2	5.3	4	3.6
Intercept Difference										
Over	1	3.4	0	0.0	0	0.0	1	2.6	2	1.8
Under	0	0.0	0	0.0	1	6.3	0	0.0	1	0.9

Over	Indicates over-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).
Under	Indicates under-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).

- AFQT exhibited no statistically significant evidence of differential prediction for nearly all Ratings examined with respect to predicting the probability of graduating from technical training.

Magnitude of prediction differences: d_{Mod} summary – probability of training graduation

	At AFQT Category IIIB Lower Bound (31)				At AFQT Category IIIA Lower Bound (50)			
Magnitude / Direction of Effect	WNH-BNH	WNH-HW	WNH-ANH	M-F	WNH-BNH	WNH-HW	WNH-ANH	M-F
Large overprediction ($d_{Mod} \geq .80$)	1							
Moderate overprediction ($.50 \leq d_{Mod} < .80$)					1			
Small overprediction ($.20 \leq d_{Mod} < .50$)				1				1
Very small overprediction ($0 \leq d_{Mod} < .20$)					1			1
Very small underprediction ($-.20 < d_{Mod} < 0$)							1	1
Small underprediction ($-.50 < d_{Mod} \leq -.20$)	1						1	
Moderate underprediction ($-.80 < d_{Mod} \leq -.50$)			1	2				
Large underprediction ($d_{Mod} \leq -.80$)			1					
Totals	2	0	2	3	2	0	2	3

- Table shows the magnitude of differences in prediction for Ratings that showed statistically significant evidence of differential prediction for AFQT scores.
- Cells show number of AFS that exhibited differences in prediction of the given d_{Mod} size at AFQT Category IIIB lower bound (31) or AFQT Category IIIA (50).
- Most prediction differences were **small or very small** at both of these AFQT score levels.

Differential Prediction Results
Attrition

A caveat about examining attrition as an outcome

- We examined 36-month attrition as a criterion (even though the AFQT was not designed to predict attrition).
- Consider this an exploratory examination – the results provide a point of comparison for future differential prediction research involving selection and classification measure more targeted at predicting attrition (e.g., TAPAS).

Differential prediction outcome summary – probability of 36-month attrition

	Differential Prediction Outcome			
Service	WNH-BNH	WNH-HW	WNH-ANH	M-F
Air Force	Intercept Difference	Intercept Difference	Slope Difference	Intercept Difference
Army	Slope Difference	Slope Difference	Slope Difference	Intercept Difference
Coast Guard	No Difference	No Difference	No Difference	No Difference
Marine Corps	Slope Difference	Slope Difference	Intercept Difference	Slope Difference
Navy	No Difference	Slope Difference	Slope Difference	Intercept Difference

- AFQT exhibited statistically significant evidence of differential prediction for predicting 36-month attrition for nearly all Service x subgroup contrast combinations examined. Exceptions were for the WNH-BNH contrast in the Navy and all subgroup contrasts in the Coast Guard.

Magnitude of prediction differences – probability of 36-month attrition

	At AFQT Category IIIB Lower Bound (31)				At AFQT Category IIIA Lower Bound (50)			
Service	WNH-BNH	WNH-HW	WNH-ANH	M-F	WNH-BNH	WNH-HW	WNH-ANH	M-F
Air Force	-.013	.062	.125	-.035	-.011	.047	.092	-.027
Army	.068	.154	.190	-.083	.037	.105	.144	-.079
Coast Guard								
Marine Corps	-.009	.089	.088	-.038	-.019	.062	.049	-.054
Navy		.088	.149	-.022		.070	.097	-.024

- In contrast to the previous sections, the table above shows raw differences in predicted probabilities of 36-month attrition from subgroup-specific regression lines for Service x subgroup contrast combinations that showed statistically significant evidence of differential prediction for AFQT scores.
- Positive values indicate a common regression line would over-predict the given focal group's (BNH, HW, ANH, F) probability of 36-month attrition for applicants at the given AFQT score relative to a subgroup-specific regression line, and negative values indicate a common regression line would under-predict the given focal group's probability of 36-month attrition for applicants at the given AFQT score relative to a subgroup-specific regression line.
- Relatively large amounts of over-prediction of probabilities of 36-month attrition were evident for Hispanic White and Asian non-Hispanic Servicemembers (particularly within the Army).

Summary, Future Research, and Questions for the DAC

Summary

Service-Specific Training and Job Knowledge Criteria

- Across Services, we examined 664 military occupation/training course-by-subgroup contrast combinations for evidence of differential prediction for the AFQT for predicting Service-specific criteria. Most combinations examined (79.1%) yielded no statistically significant evidence of differential prediction.
- Consistent with past research in the civilian and military cognitive ability testing literature, use of a common regression line for the AFQT tended to over-predict performance for Black individuals and under-predict performance for White individuals relative to use of subgroup-specific regression lines (e.g., Berry, 2015, Wise et al., 1992). Findings for other focal subgroups reflected a mix of over- and under-prediction when differences were found.
- Across Services, of the 139 military occupation/training course-by-subgroup contrast combinations that exhibited statistically significant evidence of differential prediction for the AFQT, most exhibited **small or very small** prediction differences at AFQT Category IIIB lower bound (31) and AFQT Category IIIA lower bound (50).

Attrition Criterion

- Relative to the Service-specific criteria, there was more evidence of differential prediction of the AFQT for predicting probabilities of 36-month attrition (caveat – AFQT was not designed to predict attrition).
- Prediction differences for attrition were particularly strong for Hispanic White and Asian non-Hispanic Servicemembers relative to White non-Hispanic Servicemembers, with evidence that a common regression line would result in relatively large over-prediction of attrition for Hispanic White and Asian non-Hispanic Servicemembers (particularly in the Army).
  - Differences suggest another reason why it is important to consider non-cognitive measures (e.g., TAPAS) if one's objective is to predict first term attrition with a selection/classification measure.

Future Research

Caveat: Ideal things to do, but all may not be feasible or practical to do in reality…

- **Follow-up on instances of large over/under prediction:** Follow-up on those occupations/training courses where there was evidence of (a) large over-/under-prediction of criteria, and (b) relatively larger sample sizes. Can omitted variables explain the presence of over-/under-prediction (e.g., considering education tier or TAPAS for attrition)?
- **Explore potential solutions for limited statistical power for many occupations/courses:** We used five years' worth of data, yet still faced limited power statistical power for many occupations/courses. Alternatives here may require a shift away from statistical significance testing via the Cleary approach or use of such an approach coupled with robust/defensible occupation/course clustering methods.
- **Gather data on stronger criteria:** Navy and Marine Course training graduation criteria suffered from very high graduation rates (limiting variance to be predicted), Air Force awarding course grades were also limited in variance. Job knowledge tests in they Army serve as a proximal determinant of job performance (i.e., declarative knowledge), but in and of themselves are not performance criteria. The "criterion problem" is an age-old problem that is not limited to DoD, but efforts should continue to be made to improve criterion measurement.
- **Consider differential prediction for Service-specific ASVAB composites used for occupational qualification**. The focus here was on differential prediction for AFQT, but other Service-specific ASVAB composites are used to determine whether an applicant qualifies for a given occupation. Future research might perform differential prediction analyses like ones conducted in this study, only focusing on Service-specific ASVAB composites instead of AFQT.

Questions for the DAC

1. Any specific thoughts on our "modified" Clearly approach (e.g., concerns, ways to improve)?

2. Any specific thoughts on other factors we should examine that may explain the larger instances of over/under prediction?

3. Any specific thoughts on approaches for dealing with the limited statistical power we observed for low *N* occupations (e.g., specific factors on which to cluster occupations, other methods)?

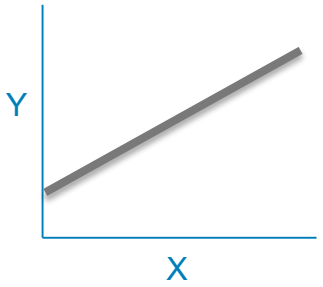
Other Questions?

Back Up Slides Part 1: Primer on Evaluating Test Scores for Differential Prediction

Visualizing differences in prediction

No Difference

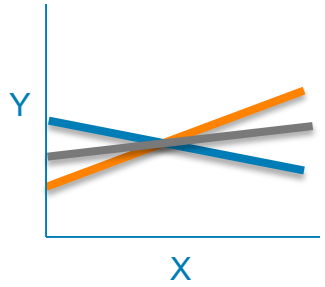
At a common test score, criterion Y is predicted to be the same for members of different subgroups.

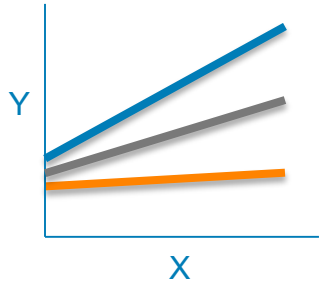


A common regression line accounts for the relationship between the test (X) and criterion (Y) in both subgroups examined.

Slope Difference

Subgroup-specific regression lines differ in slope.

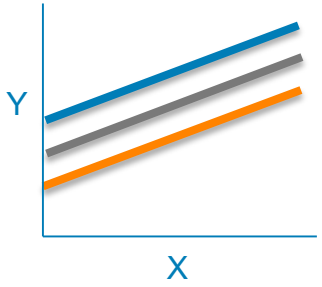




Use of a common regression line may result in over-prediction of criterion Y for one subgroup and under-prediction of criterion Y for the other subgroup (relative to subgroup-specific regression lines), but direction and extent of differences depends on the test score considered.

Intercept Difference

Subgroup-specific regression lines differ in intercept, but not slope.



Use of a common regression line results in over-prediction of criterion Y for the subgroup represented by the orange line and under-prediction of criterion Y for the subgroup represented by the blue line.

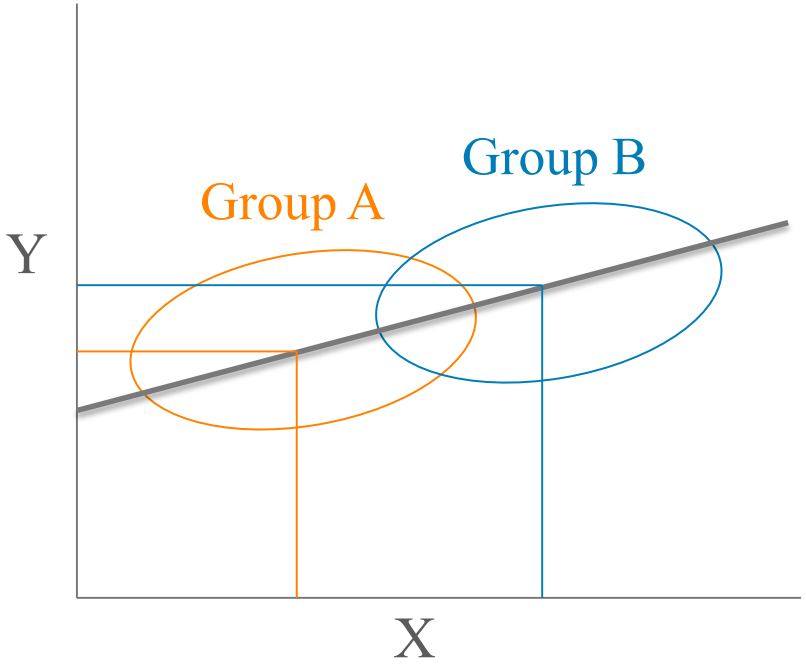
Common regression line Subgroup 1 regression line Subgroup 2 regression line

Important note about mean score differences and differential prediction

- Groups may exhibit mean score differences on a test (e.g., AFQT), but that does not necessarily mean scores from that test exhibit differential prediction.

- In the example below, the groups being compared differ in terms of their mean X (test) and mean Y (criterion) scores, but X exhibits no evidence of differential prediction, as a common regression line accounts for X-Y relations.

- **This point is critical to keep in mind**. Historically, there have been mean score differences between White and Black test-takers on cognitive aptitude tests such as the AFQT, but that does not necessarily mean scores on those tests exhibit differential prediction.



Evaluating scores for differential prediction

- The de facto approach for evaluating differential prediction in the field of I-O Psychology involves a three-step comparison of nested regression models – sometimes referred to as the Cleary approach:

1. $y = b_{10} + b_{11} * AFQT + e$
2. $y = b_{20} + b_{21} * AFQT + b_{22} * Subgroup + e$
3. $y = b_{30} + b_{31} * AFQT + b_{32} * Subgroup + b_{33} * (AFQT * Subgroup) + e$

- Evaluations of differential prediction involve testing whether predictor-criterion relations within subgroups of interest can be accounted for by a common regression line and, if not, clarifying the nature of the differences between regression lines for the said subgroups.

Evaluating scores for differential prediction

- The de facto approach for evaluating differential prediction in the field of I-O Psychology involves a three-step comparison of nested regression models – sometimes referred to as the Cleary approach:

1. $y = b_{10} + b_{11} * AFQT + e$
2. $y = b_{20} + b_{21} * AFQT + b_{22} * Subgroup + e$
3. $y = b_{30} + b_{31} * AFQT + b_{32} * Subgroup + b_{33} * (AFQT * Subgroup) + e$

The first step involves comparing Models 1 and 3.

If the increment in R^2 of Model 3 over Model 1 is statistically significant, it suggests the presence of differential prediction, but does not clarify whether this is due to intercept or slope differences between subgroups.

If the increment in R^2 of Model 3 over Model 1 is *not* significant, the process stops.

Evaluating scores for differential prediction

- The de facto approach for evaluating differential prediction in the field of I-O Psychology involves a three-step comparison of nested regression models – sometimes referred to as the Cleary approach:

1. $y = b_{10} + b_{11} * AFQT + e$
2. $y = b_{20} + b_{21} * AFQT + b_{22} * Subgroup + e$
3. $y = b_{30} + b_{31} * AFQT + b_{32} * Subgroup + b_{33} * (AFQT * Subgroup) + e$

If the increment in R^2 of Model 3 over Model 1 was significant, the second step involves comparing Models 2 and 3.

If the increment in R^2 of Model 3 over Model 2 is statistically significant, it indicates that the subgroups' regression lines exhibit slope differences.

Evaluating scores for differential prediction

- The de facto approach for evaluating differential prediction in the field of I-O Psychology involves a three-step comparison of nested regression models – sometimes referred to as the Cleary approach:

1. $y = b_{10} + b_{11} * AFQT + e$
2. $y = b_{20} + b_{21} * AFQT + b_{22} * Subgroup + e$
3. $y = b_{30} + b_{31} * AFQT + b_{32} * Subgroup + b_{33} * (AFQT * Subgroup) + e$

If the increment in R^2 of Model 3 over Model 2 was *not* significant (i.e., no evidence of slope differences), the third step involves comparing Models 1 and 2.

If the increment in R^2 of Model 2 over Model 1 is statistically significant, it indicates that the subgroups' regression lines exhibit intercept differences.

Back Up Slides Part 2: AFQT and Criterion Descriptives and Subgroup Differences

Composition of applicant population

Variable	*n*	%	Variable	*n*	%
FY AFQT Completed			Gender		
FY14	311,464	19.4	Male	1,218,211	76.0
FY15	318,261	19.8	Female	385,523	24.0
FY16	316,737	19.7	Missing	15	0.0
FY17	319,779	19.9			
FY18	337,508	21.0	Race/Ethnicity		
			WNH	813,193	50.7
Service Applied To			BNH	344,067	21.5
Air Force	256,238	16.0	HW	271,504	16.9
Army	791,282	49.3	ANH	77,265	4.8
Coast Guard	31,704	2.0	Other/Missing	97,720	6.1
Marine Corps	238,073	14.8			
Navy	286,390	17.9	Gender-Race/Ethnicity		
Missing	62	0.0	M-WNH	657,779	41.0
			M-BNH	227,969	14.2
Service Accessed To			M-HW	204,260	12.7
Air Force	131,833	8.2	M-ANH	58,932	3.7
Army	460,290	28.7	F-WNH	155,413	9.7
Coast Guard	15,435	1.0	F-BNH	116,095	7.2
Marine Corps	164,429	10.3	F-HW	67,244	4.2
Navy	167,486	10.4	F-ANW	18,332	1.1
Did Not Access	664,276	41.4	Other/Missing	97,725	6.1
AFQT Category					
I	93,499	5.8			
II	519,198	32.4			
IIIA	372,363	23.2			
IIIB	418,540	26.1			
IV	166,579	10.4			
V	33,570	2.1			

Note. Total *N* = 1,603,749.

AFQT Categories reflect AFQT score ranges defined as follows: I (93-99), II (65-92), IIIA (50-64), IIIB (31-49), IV (10-30), and V (1-9).

WNH = White non-Hispanic; HW = Hispanic White; BNH = Black non-Hispanic; ANH = Asian non-Hispanic; M = Male; F = Female.

AFQT scores by subgroup in applicant population

Sample	*n*	*M*	*SD*	*d*
Overall	1,603,749	56.62	22.84	-
Gender				
Male	1,218,211	58.30	22.77	-
Female	385,523	51.30	22.26	0.31
Race/Ethnicity				
WNH	813,193	62.99	21.30	-
BNH	344,067	45.24	21.29	0.83
HW	271,504	51.33	21.92	0.54
ANH	77,265	58.50	24.64	0.20
Gender-Race/Ethnicity				
M-WNH	657,779	63.95	21.23	-
M-BNH	227,969	46.53	21.59	0.82
M-HW	204,260	52.68	21.98	0.52
M-ANH	58,932	59.70	24.69	0.19
F-WNH	155,413	58.89	21.15	0.24
F-BNH	116,095	42.72	20.46	1.01
F-HW	67,244	47.21	21.21	0.79
F-ANW	18,332	54.66	24.09	0.41

Note. *d* = Standardized mean difference between referent subgroups (e.g., Male, White non-Hispanic, Male White non-Hispanic) and focal subgroups groups (e.g., Female, Black non-Hispanic, Male Black non-Hispanic) on AFQT.

$d = (M_{referent} - M_{focal}) / SD_{Pooled}$, where the pooled *SD* was calculated across the given referent and focal subgroups compared.

WNH = White non-Hispanic; HW = Hispanic White; BNH = Black non-Hispanic; ANH = Asian non-Hispanic; M = Male; F = Female.

Summary of subgroup mean differences on AFS awarding course grades

	$d_{M\text{-}F}$	$d_{WNH\text{-}BNH}$	$d_{WNH\text{-}HW}$	$d_{WNH\text{-}ANH}$
k	50	54	51	10
Mean	-0.02	0.23	0.08	-0.13
SD	0.16	0.19	0.12	0.20
Skew	-0.35	-0.13	0.02	-0.15
Max	0.26	0.60	0.32	0.17
Min	-0.38	-0.25	-0.20	-0.50
Percentile				
95	0.21	0.53	0.27	0.13
75	0.10	0.36	0.15	0.04
50	-0.00	0.24	0.07	-0.18
25	-0.14	0.10	0.01	-0.23
5	-0.31	-0.05	-0.08	-0.41

Note. *d* = Standardized mean difference between referent subgroups (e.g., Male, White non-Hispanic, Male White non-Hispanic) and focal subgroups (e.g., Female, Black non-Hispanic, Male Black non-Hispanic) on awarding course grades.

$d = (M_{referent} - M_{focal}) / SD_{Pooled}$, where the pooled *SD* was calculated across the given referent and focal subgroups compared.

M = Male. F = Female. WNH = White non-Hispanic. HW = Hispanic White. BNH = Black non-Hispanic. ANH = Asian non-Hispanic.

- Table provides summary of standardized subgroup mean differences on **AFS awarding course grades**.
- *k* = # of AFS for which awarding course grades were compared for the given subgroup pair.
- Reported statistics are across AFS.

Summary of subgroup mean differences on Army-wide job knowledge test scores

	$d_{M\text{-}F}$	$d_{WNH\text{-}BNH}$	$d_{WNH\text{-}HW}$	$d_{WNH\text{-}ANH}$
k	20	22	28	11
Mean	0.09	0.44	0.21	0.14
SD	0.16	0.14	0.14	0.13
Skew	-0.02	-0.99	0.05	-0.19
Max	0.39	0.66	0.45	0.32
Min	-0.26	0.01	-0.04	-0.06
Percentile				
95	0.32	0.63	0.42	0.30
75	0.18	0.52	0.30	0.24
50	0.10	0.43	0.21	0.16
25	-0.02	0.39	0.10	0.02
5	-0.09	0.26	-0.01	-0.03

Note. *d* = Standardized mean difference between referent subgroups (e.g., Male, White non-Hispanic, Male White non-Hispanic) and focal subgroups (e.g., Female, Black non-Hispanic, Male Black non-Hispanic) on Army-wide JKT scores.

$d = (M_{referent} - M_{focal}) / SD_{Pooled}$, where the pooled *SD* was calculated across the given referent and focal subgroups compared.

M = Male. F = Female. WNH = White non-Hispanic. HW = Hispanic White. BNH = Black non-Hispanic. ANH = Asian non-Hispanic.

- Table provides summary of subgroup mean differences on **Army-wide JKT scores**.
- *k* = # of MOS for which Army-wide JKT scores were compared for the given subgroup pair.
- Reported statistics are across MOS.

Summary of subgroup mean differences on MOS-specific job knowledge test scores

MOS		$d_{M\text{-}F}$	$d_{WNH\text{-}BNH}$	$d_{WNH\text{-}HW}$	$d_{WNH\text{-}ANH}$
11B	Infantryman		0.56	0.42	0.19
12B	Combat Engineer	0.54	0.93	0.56	0.30
13D	Field Artillery Rocket Crewman	0.41	0.45	0.16	
13F	Joint Fire Support Specialist		0.50	0.24	
19D	Cavalry Scout		0.46	0.21	0.12
19K	M1 Armor Crewman		0.33	-0.01	
31B	Military Police	0.34	0.47	0.36	0.00
68W	Combat Medic Specialist	-0.19	0.37	0.20	0.30
88M	Motor Transport Operator	0.06	0.65	0.46	0.54
91B	Wheeled Vehicle Mechanic	0.45	0.62	0.34	0.32

Note. *d* = Standardized mean difference between referent subgroups (e.g., Male, White non-Hispanic, Male White non-Hispanic) and focal subgroups (e.g., Female, Black non-Hispanic, Male Black non-Hispanic) on MOS-specific JKT scores.

$d = (M_{referent} - M_{focal}) / SD_{Pooled}$, where the pooled *SD* was calculated across the given referent and focal subgroups compared.

M = Male. F = Female. WNH = White non-Hispanic. HW = Hispanic White. BNH = Black non-Hispanic. ANH = Asian non-Hispanic.

- Table provides summary of subgroup mean differences on **MOS-specific JKT scores**.
- Missing values indicate insufficient data were available to calculate subgroup mean differences for the given MOS.

Summary of Marine Corps training course graduation rates by subgroup

	Gender		Race/Ethnicity			
	M	F	WNH	BNH	HW	ANH
k	68	68	122	64	121	26
Mean	.92	.91	.92	.86	.91	.94
SD	.12	.14	.12	.15	.14	.06
Skew	-3.82	-2.90	-3.83	-2.84	-3.96	-1.52
Max	1.00	1.00	1.00	1.00	1.00	1.00
Min	.31	.30	.17	.18	.09	.74
Percentile						
95	.99	1.00	.99	.99	.99	.99
75	.98	.98	.98	.98	.98	.98
50	.96	.95	.95	.94	.96	.96
25	.92	.91	.90	.85	.90	.90
5	.81	.60	.77	.66	.78	.74

Note. M = Male. F = Female. WNH = White non-Hispanic. HW = Hispanic White. BNH = Black non-Hispanic. ANH = Asian non-Hispanic.

- Table provides summary of **training course graduation rates** by subgroup:
  - 1 = Graduation without setback
  - 0 = Graduation with setback or non-graduate
- *k* = # of training courses for which graduation status was calculated for given subgroup.
- Reported statistics are across training courses.

Summary of Navy training graduation rates by subgroup

		Gender		Race/Ethnicity			
	Overall	M	F	WNH	BNH	HW	ANH
k	45	38	38	37	29	28	16
Mean	.94	.94	.94	.93	.91	.92	.96
SD	.08	.09	.08	.08	.10	.10	.04
Skew	-2.66	-2.89	-2.95	-2.89	-2.05	-2.60	-1.47
Max	1.00	1.00	1.00	1.00	1.00	.99	.99
Min	.56	.55	.58	.55	.55	.51	.86
Percentile							
95	1.00	1.00	.99	1.00	.99	.99	.99
75	.99	.99	.99	.98	.97	.98	.98
50	.97	.97	.97	.95	.94	.96	.97
25	.92	.93	.92	.92	.90	.91	.95
5	.80	.76	.84	.81	.74	.74	.90

Note. M = Male. F = Female. WNH = White non-Hispanic. HW = Hispanic White. BNH = Black non-Hispanic. ANH = Asian non-Hispanic.

- Table provides summary of **training graduation rates** by subgroup:
  - 1 = Graduation without setback
  - 0 = Graduation with setback or non-graduate
- *k* = # of Ratings for which graduation status was calculated for given subgroup.
- Reported statistics are across Ratings.

36-month attrition rates by Service and subgroup

	Proportion Attrit through 36-months of Service									
	Air Force		Army		Coast Guard		Marine Corps		Navy	
Sample	*n*	*p*	*n*	*p*	*n*	*p*	*n*	*p*	*n*	*p*
Race/Ethnicity										
WNH	59,124	.14	109,829	.28	8,183	.15	62,203	.16	63,418	.23
BNH	16,799	.16	46,525	.29	691	.20	10,235	.19	21,442	.25
HW	14,847	.11	35,348	.22	1,818	.15	26,348	.12	13,722	.18
ANH	4,179	.08	10,989	.17	211	.12	2,750	.11	5,222	.14
Gender										
Male	79,303	.13	172,199	.25	10,104	.16	95,917	.14	92,582	.21
Female	22,840	.16	34,446	.34	1,574	.17	10,691	.21	30,817	.25

Note. *n* = # of Regular component NPS accessions of the given race/ethnicity or gender for which the attrition criterion was calculated. *p* = Proportion attrit through 36-months of service

Back Up Slides Part 3: Summary Details

Differential prediction outcome summary: Service-specific criteria

Across Services, we examined 664 military occupation/training course-by-subgroup contrast combinations for evidence of differential prediction for the AFQT for predicting Service-specific criteria. Most combinations examined (79.1%) yielded no statistically significant evidence of differential prediction.

	WNH-BNH		WNH-HW		WNH-ANH		M-F		Grand Totals	
# of Combos	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Total	178		236		70		180		664	
No Difference	121	68.0	203	86.0	54	77.1	147	81.7	525	79.1
Slope Difference										
Over	3	1.7	2	0.8	0	0.0	3	1.7	8	1.2
Under	9	5.1	9	3.8	7	10.0	7	3.9	32	4.8
Intercept Difference										
Over	41	23.0	16	6.8	5	7.1	11	6.1	73	11.0
Under	4	2.2	6	2.5	4	5.7	12	6.7	26	3.9

Over	Indicates over-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).
Under	Indicates under-prediction of criterion for focal subgroup at the lower bound of AFQT Cat IIIB (31).

Consistent with past research in the civilian and military cognitive ability testing literature, use of a common regression line for the AFQT tended to over-predict performance for Black individuals and under-predict performance for White individuals relative to use of subgroup-specific regression lines (e.g., Berry, 2015, Wise et al., 1992).

Findings for other focal subgroups reflected a mix of over- and under-prediction when differences were found.

Magnitude of prediction differences: d_{Mod} summary for Service-specific criteria

Across Services, of the 139 military occupation/training course-by-subgroup contrast combinations that exhibited statistically significant evidence of differential prediction for the AFQT, most exhibited **small or very small** prediction differences at AFQT Category IIIB lower bound (31) and AFQT Category IIIA (50).

	At AFQT Category IIIB Lower Bound (31)				At AFQT Category IIIA Lower Bound (50)			
Magnitude / Direction of Effect	WNH-BNH	WNH-HW	WNH-ANH	M-F	WNH-BNH	WNH-HW	WNH-ANH	M-F
Large overprediction (dMod ≥ .80)	6	2	0	3	2	2	0	3
Moderate overprediction (.50 ≤ dMod < .80)	6	3	0	0	9	1	0	0
Small overprediction (.20 ≤ dMod < .50)	17	4	4	8	31	7	5	6
Very small overprediction (0 ≤ dMod < .20)	15	9	1	3	11	10	2	6
Very small underprediction (-.20 < dMod < 0)	6	5	3	1	1	9	4	6
Small underprediction (-.50 < dMod ≤ -.20)	3	4	2	10	2	2	2	10
Moderate underprediction (-.80 < dMod ≤ -.50)	3	3	4	6	0	1	3	1
Large underprediction (dMod ≤ -.80)	1	3	2	2	1	1	0	1
Totals	57	33	16	33	57	33	16	33

Differential prediction of AFQT scores for predicting attrition

- Relative to the Service-specific criteria, there was more evidence of differential prediction of the AFQT for predicting probabilities of 36-month attrition.
- Prediction differences were particularly strong for Hispanic White and Asian non-Hispanic Servicemembers relative to White non-Hispanic Servicemembers, with evidence that a common regression line would result in relatively large over-prediction for Hispanic White and Asian non-Hispanic Servicemembers (particularly in the Army).
    - Note, unlike the positively valenced Service-specific criteria examined, over-prediction of attrition disadvantages the subgroup being over-predicted.
- When prediction differences were found for Black non-Hispanic and Female Servicemembers, they typically indicated a common regression line would result in under-prediction of probabilities of attrition
    - Exception was for Black non-Hispanic Soldiers in the Army – which indicated the over-prediction of attrition probabilities.