Damos Aviation Services

Meeting pilot hiring needs.

How Do We Reduce Bias in Aviation Selection?

Diane L. Damos, Ph.D.

Damos Aviation Services, Inc.

Overview

- Discuss adverse impact on women during selection for training or hiring as pilots
- Describe methods for identifying the source of the adverse impact
- Suggest ways to decrease adverse impact when assessing abilities known for large male/female differences
- Identify abilities with potentially little adverse impact

Literature Review

- Reviewed pilot selection studies reporting batteries
  - 1996 to present
  - Discuss tests in detail
  - 13 Western European, British, and US studies
    - US and British studies are military
    - Western European are mainly civilian

Where are the Adverse Impact Data?

- No European or British data
- No US civilian data
- US Air Force
  - Air Force Officer Qualifying Test (AFOQT)
  - Apparatus tests

Discrimination

- Definition—failure to treat all persons equally when no reasonable distinction can be found between those favored and those not favored (Black's Law Dictionary)
  - Deliberate

Disparate (Adverse) Impact

- Definition—substantially different rate of selection in hiring that works to the disadvantage of members of a race, sex or ethnic group. It is an unwanted or unanticipated repercussion caused by a specific practice
  - Accidental

Bias and Fair

- Bias—psychometric properties of the test
- Fair—judgement, may be based on adverse impact

Interpreting Results

- Cohen's d
  - Measure of effect size
  - Dimensionless
  - Correction for unequal n's
  - Effect sizes
    - d = 0.2 small
    - d = 0.5 medium
    - d = 0.8 large (half of areas do not overlap)

Structure of the Talk

- High altitude
  - Many tests in the battery show large d's
    - How to interpret this
      - ✓ Battery level
      - ✓ Test level
- Low altitude
  - Which categories of tests show bias?
  - Which categories of tests show little or no bias?

Background

- US Air Force AFOQT
  - Updated ≈ every 7 years
  - General intelligence test/academic
  - Version T is current
    - 16 tests
    - 5 composites—Verbal, Quantitative, Academic (Verbal + Quantitative), Pilot, Navigator-Technical

Background

- US Air Force Apparatus tests
  - Intermittent use
    - 1942-1955
    - BAT—1992 to 2006
    - TBAS—2006 to present day
      - ✓ Hand tracking
      - ✓ Foot tracking
      - ✓ Timesharing

Carretta (1997)

- Why do so few women pass the pilot selection process?
- AFOQT Means
  - Sample—Officer candidates
    - Male = 219, 887
    - Females = 50, 081
  - Mean differences?
    - Males > females on 15 of 16 AFOQT tests
    - No difference on Verbal Analogies
    - Composites ranged from d = 0.08 (verbal) to 0.69 (pilot)

Carretta (1997)

- What is the cause? Psychometric problem? Real difference? Something else?
  - Battery level
    - Confirmatory factor analyses by gender. Want to see:
      - ✓ Same number and identity of factors
      - ✓ Factors account for the same proportions of total and common variance

Carretta (1997)

- Is AFOQT factor structure the same for male versus female officer candidates
  - Factor analysis showed identical factor structure, prop of variance accounted for very similar
- Something else?
  - Correct population?

Carretta (1997)

- AFOQT Means—Pilot candidates
  - Sample—Pilot Officer candidates
    - Male = 9, 239
    - Females = 237
  - Mean differences?
    - Males > females on 6 of 16 AFOQT tests
    - Mean difference d = 0.08
    - Composites ranged from d = -0.48 (verbal) to 0.20 (navigator/technical); mean d = - 0.10
    - Male composites > female composites only on navigator/technical

Carretta (1997)

- BAT—Pilot candidates only
  - 4 tests
    - 2 psychomotor
    - STM
    - Timesharing test—tracking + STM
  - Candidates
    - Male = 4,888
    - Females = 465

Carretta (1997)

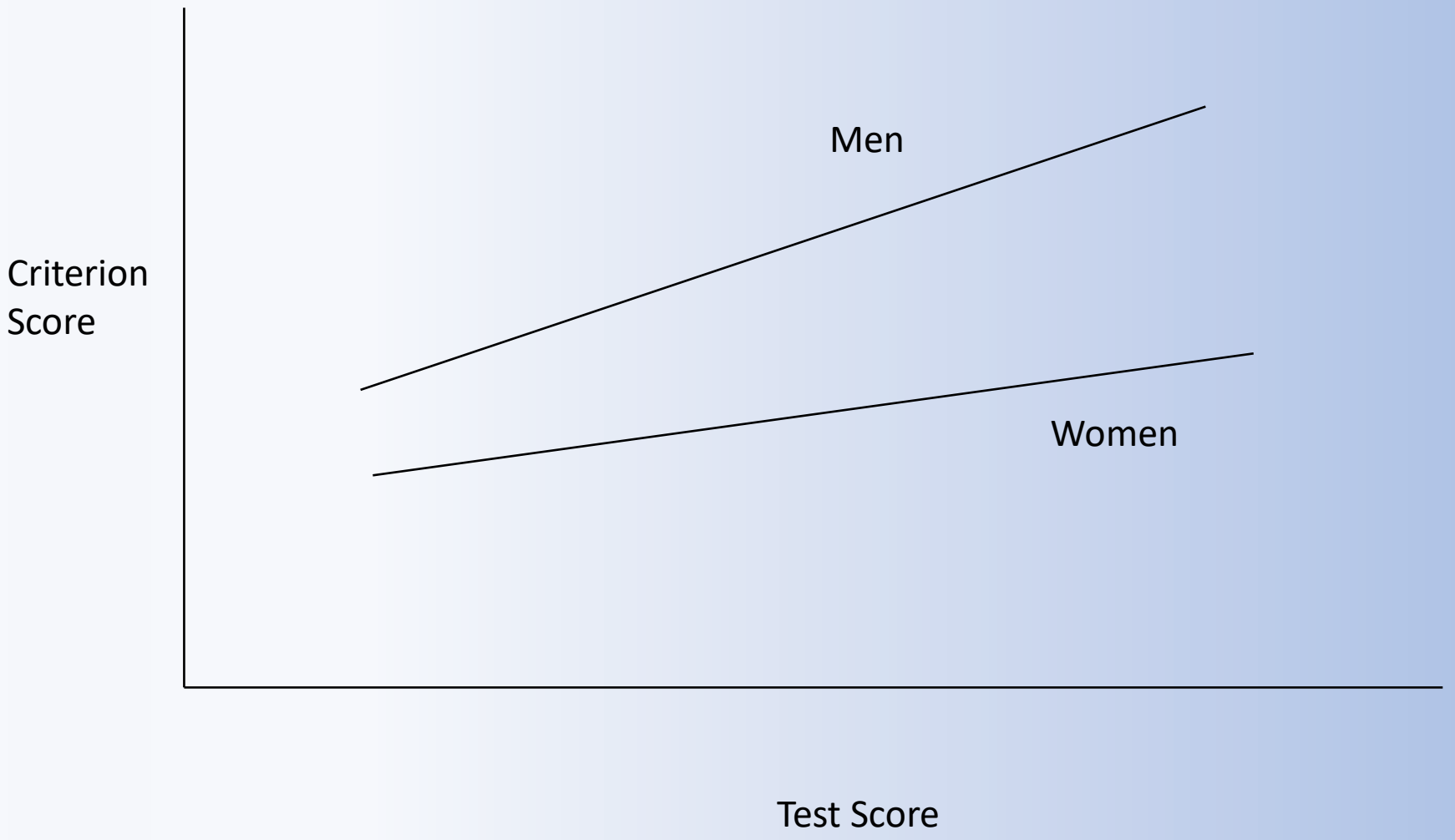
- BAT—Pilot candidates only
  - Mean differences?
    - Males > females on 4 of 4 tests
    - STM d = 0.10
    - Psychomotor composite d = 1.68
    - Timesharing composite (includes tracking) d = 1.04

Carretta (1997)

- Is BAT factor structure the same for male versus female pilot candidates?
  - Factor analysis showed identical factor structure, prop of variance accounted for very similar

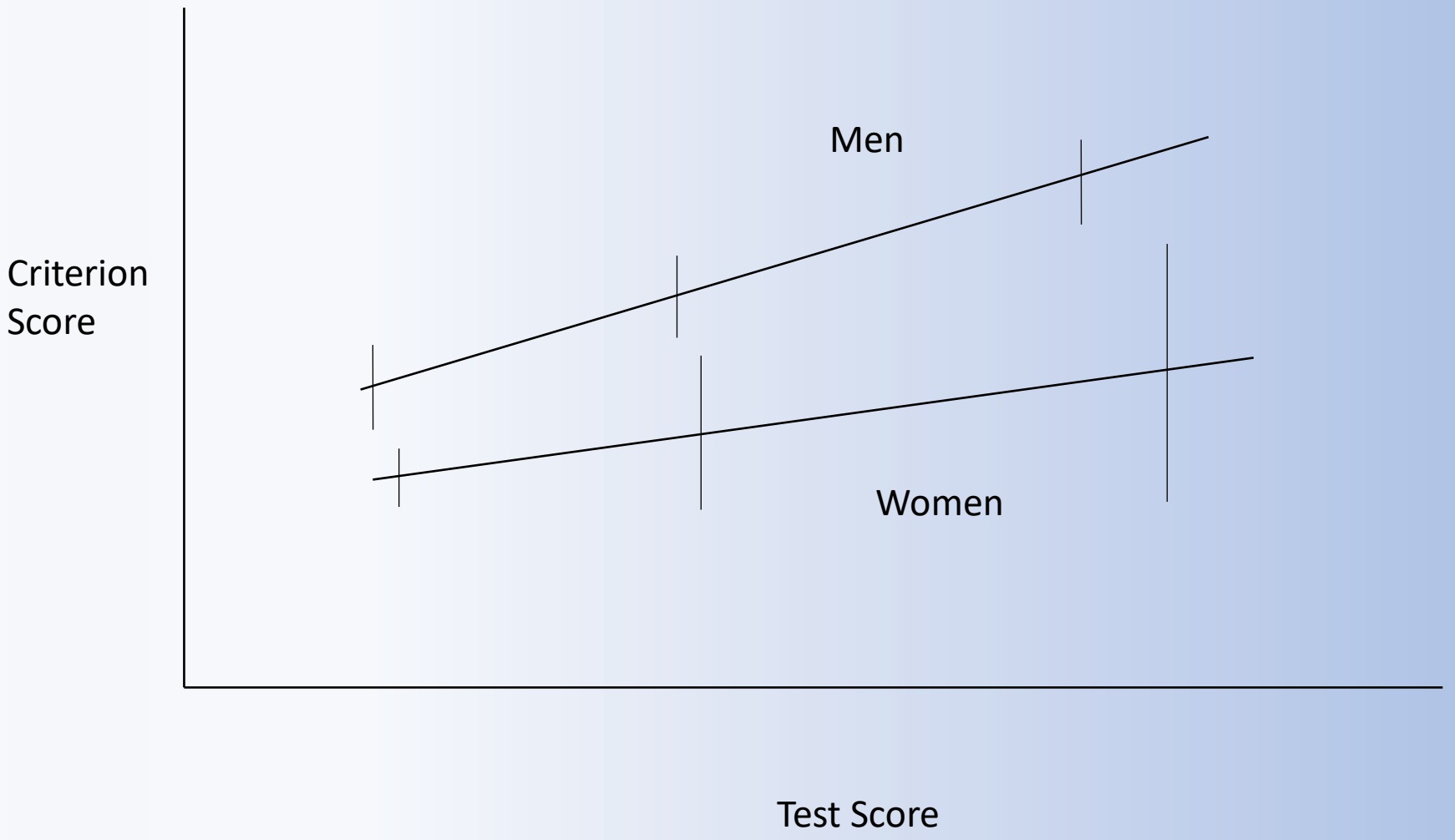
Individual Test Level

- Differential predictive validity
  - Regression for males and females separately
    - Slope differences?
      - ✓ “yes”


Criterion Score
Men
Women
Test Score

Individual Test Level

- Differential predictive validity
  - Regression for males and females separately
    - Slope differences? Yes!!!
      - ✓ Statisticians say “Do item analysis and change as necessary”
      - ✓ Works if have large samples and lots of time. Not for apparatus tests.
      - ✓ Check homoscedasticity

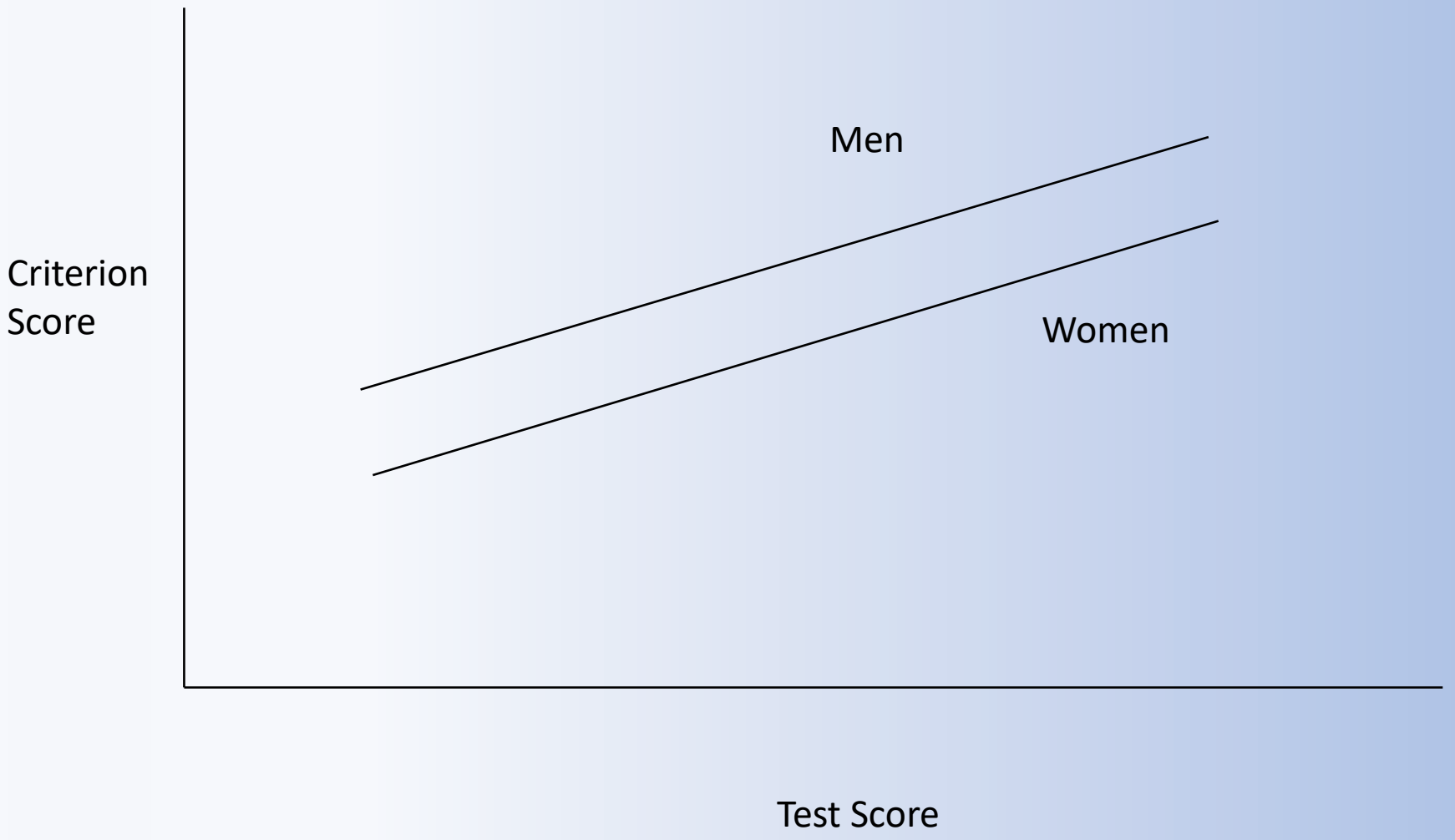

Criterion Score
Men
Women
Test Score

Individual Test Level

- Differential predictive validity
  - Regression for males and females separately
    - Slope differences? Yes!!!
      - Statistically correct for heteroscedasticity

Individual Test Level

- Differential predictive validity
  - Regression for males and females separately
    - Slope differences?
      - ✓ “No”


Men
Women
Criterion Score
Test Score

Individual Test Level

- Regression for males and females separately
  - Slope differences?
    - ✓ If “No”
      - Test is not psychometrically biased

Trent & Aguilar (2020)

- Compared male versus female pilot officer candidates
- AFOQT—Pilot composite
  - Instrument Comprehension
  - Perceptual Speed Test
  - Aviation Information
  - Quantitative
- TBAS
  - Tracking (hand, foot)
  - Timesharing (hand and foot)

Trent & Aguilar (2020)

- N = 14,214
- Male 12, 451; Female 1,763
- AFOQT pilot composite d = 0.67
- TBAS tracking results
  - Hand d = 1.52
  - Timesharing scores
    - Hand d = 1.45
    - Foot d = 0.32

Trent & Aguilar (2020)

- Predictive validity
  - Male 6304; Female 540

Score	R with specialized Primary training
PCSM Score	
Female	0.383
Male	0.382
AFOQT Pilot	
Female	0.354
Male	0.351

Trent & Aguilar (2020)

- Batteries—No differences
- Predicative validity is the same
- Conclusion: psychometrically sound, but..
- Still have large male-female differences
- Real differences? Something else?
- What to do?

Structure of the Talk

- High altitude
- Low altitude
  - Tests of which abilities show large male-female differences?
  - Tests of which abilities show small or no differences?

What Tests Are We Using?

- Most common tests (13 batteries)
  - Psychomotor (hands)—10 ★
  - Spatial—10 ★
  - Quantitative—8
  - Personality—7
  - Perceptual speed—6
  - Multiple-task (timesharing)—5

Category—Psychomotor

- What are the problems with these tests?
- What are we testing?
  - Eye-hand coordination
  - Eye-hand-foot coordination
  - Tracking ability
- Fleishman taxonomy
  - Multi-limb coordination
  - Precision control
  - Rate control
  - Response orientation
- No eye-hand coordination ability, no tracking ability
- What is being tested?

Category—Psychomotor

- Carretta (1997) psychomotor composite d = 1.68 (hand, hand)
- Trent & Aguilar (2020)
  - Hand tracking d = 1.52
  - Timesharing scores
    - Hand d = 1.45
    - Foot d = 0.32
- Historical data?
- Damos' calculations (Melton, 1947) WASP
  - Two-hand coordination (hand) d = 1.08
  - Rudder control (feet) d = -0.82

Category—Spatial

- Men are better than women, but….
  - Mental rotation (men v women, pilots vs non pilots) (Verde et al.,2013). Matched on age
  - Men faster than women
    - Pilots faster than nonpilots
    - No significant difference between male and female pilots

Category—Spatial

- Three factors (Carroll, 1993)
  - Spatial Relations
  - Spatial Orientation
  - Spatial Visualization
- D'Oliveira (2004, Study 1)
  - Men better than women on Spatial Relations
  - Not different on Spatial Visualization or dynamic spatial ability
- Fourth factor? Dynamic spatial ability (D'Oliveira, 2004)
  - Not enough data yet

Promising—Perceptual Speed

- Neglected topic. Very little gender research
- Literature confusing. Why?
  - Ackerman,Beier & Boyle (2002) 4 different types
    - Pattern matching—recognition of simple pattern
    - Scanning—scanning, comparison, and lookup
    - Memory—STM demands (digit/symbol)
    - Complex—increased memory load, scanning, and perhaps some spatial

Promising—Perceptual Speed

- Gender research
  - Ackerman, Kanfer & Goff (1995) just $p<.05$ for complex
  - Damos & Gould (2009) no sign difference ab initios
  - Hoermann & Damos (2019)
    - Ab initio males > females $p = .018$ on #Cor. No diff on #W
    - Licensed pilots no difference on either
  - WASPS $d = -0.25$
    - Outscored men on 10/10

Promising—Timesharing

- Carretta (1997) Psychomotor composite
  - Psychomotor composite (hand, hand) d = 1.68
  - Multiple-task composite (includes psychomotor) d = 1.04
- Trent & Aguilar (2020)
  - Hand tracking d = 1.52
  - Timesharing scores
    - Hand d = 1.45
    - Foot d = 0.32

Promising—Timesharing

- Cognitive psychology “myth:” Women are better timesharers than men
- What do they mean?
  - Media switching—one task involves media
  - Scheduling of large tasks

Promising Tests—Timesharing

- Cognitive laboratory task
  - Hirsch, Koch, Karbach (2019)
    - 48 men, 48 women
    - No significant differences on age, mental health physical health, STM capacity, intelligence. Women faster processing speeds, d = -0.54; men faster mental rotation, d = 0.58
    - Digit parity task, letter vowel/consonant
    - Single task, mixed blocks, dual-task
    - Switching time, decrements, concurrent speed and accuracy—No gender effects

Promising Tests—Timesharing

- Individual differences in fine-grained analyses
  - Response strategy
    - Damos, Smist & Bittner (1983)
    - Bruning & Manzey (2018)
  - Individual differences in decrements
    - Watson & Strayer (2010) Super taskers (90 males, 110 females; 3 v 2)

Summary

- Many batteries used in pilot selection have multiple tests with large gender effects
  - Batteries seem to be working the same for men and women
  - Tests seem to have same predictive validity
  - Source of differences on tests
    - Practice? Exposure?
    - Real differences?

Summary

- What to do?
  - Improve test selection for commonly used tests
    - Psychomotor
      - ✓ Carefully constructed to assess known attribute
      - ✓ Foot tracking?
      - ✓ Examine practice by gender effects
    - Spatial
      - ✓ Which abilities do we need to test?

Summary

- What to do?
  - Start investigating promising tests
    - Perceptual speed
      - ✓ Use complex perceptual speed tests
    - Timesharing
      - ✓ Response strategies
      - ✓ Individual differences in decrements