A Field Test of the 3-Option Multiple-Choice Test Format

Dr. Cal Hoffman
Los Angeles Sheriff's Department
Alliant International University

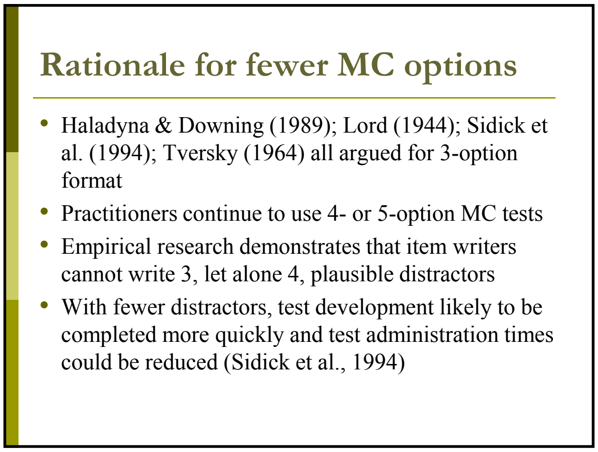
February 22, 2017

Acknowledgements

- This research relied on the contributions of Test Development Unit at LASD:
  - Carlos Valle; Chy Tashima; Daniel Kowallis; Christina Ramirez
- These tests could not have been developed without input and support of our subject matter experts
  - Too many to name
  - Around 20-25 SMEs supported each exam

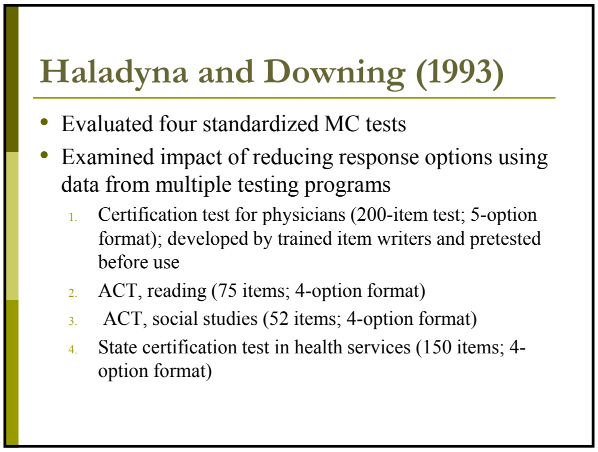
Agenda

- Rationale for move from 5- or 4- to 3-option format
- Advantages of fewer response options
- Previous research on 3-option MC test:
  - Educational
  - Applied selection
  - Meta-analyses
- Current application to sergeant promotional exam
  - Expected impact on *p*-values, item discrimination (r_{pb}), test reliability
  - *d*-values (F/M; B/W; H/W; A/W)



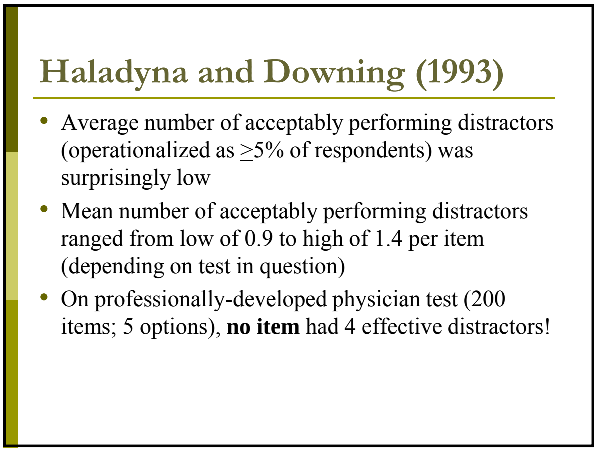
Rationale for fewer MC options

- Haladyna & Downing (1989); Lord (1944); Sidick et al. (1994); Tversky (1964) all argued for 3-option format
- Practitioners continue to use 4- or 5-option MC tests
- Empirical research demonstrates that item writers cannot write 3, let alone 4, plausible distractors
- With fewer distractors, test development likely to be completed more quickly and test administration times could be reduced (Sidick et al., 1994)



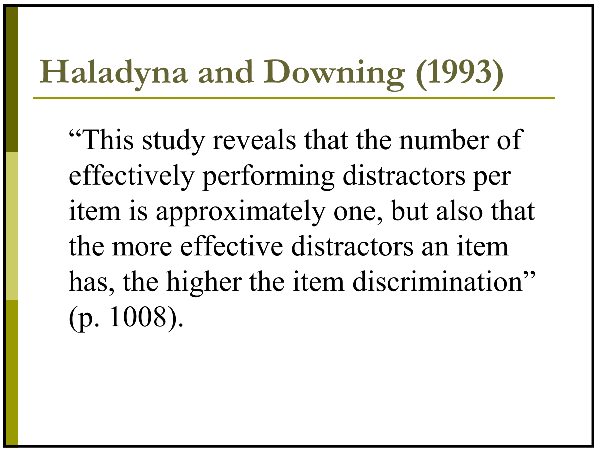
Haladyna and Downing (1993)

- Evaluated four standardized MC tests
- Examined impact of reducing response options using data from multiple testing programs
  1. Certification test for physicians (200-item test; 5-option format); developed by trained item writers and pretested before use
  2. ACT, reading (75 items; 4-option format)
  3. ACT, social studies (52 items; 4-option format)
  4. State certification test in health services (150 items; 4-option format)



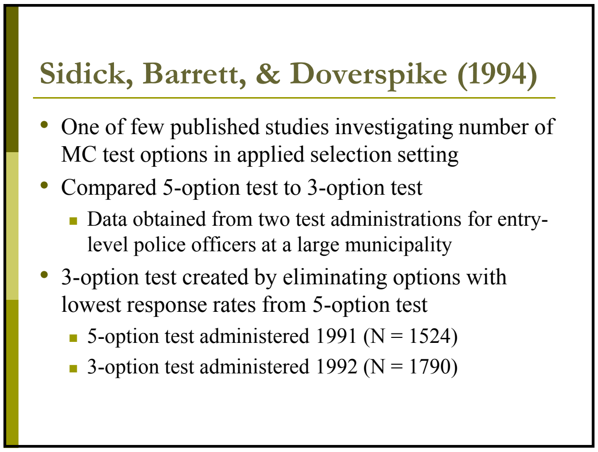
Haladyna and Downing (1993)

- Average number of acceptably performing distractors (operationalized as ≥5% of respondents) was surprisingly low
- Mean number of acceptably performing distractors ranged from low of 0.9 to high of 1.4 per item (depending on test in question)
- On professionally-developed physician test (200 items; 5 options), **no item** had 4 effective distractors!



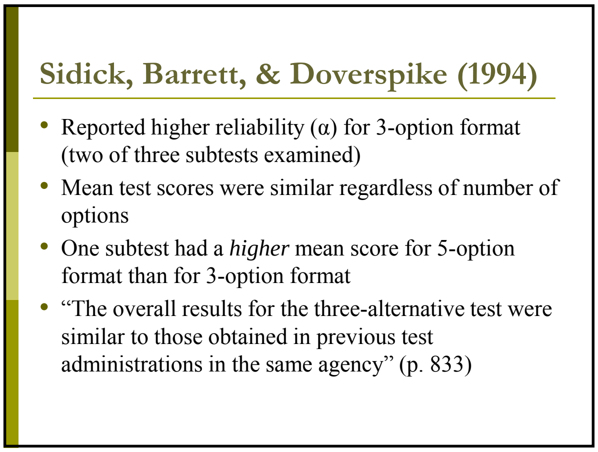
Haladyna and Downing (1993)

“This study reveals that the number of effectively performing distractors per item is approximately one, but also that the more effective distractors an item has, the higher the item discrimination” (p. 1008).



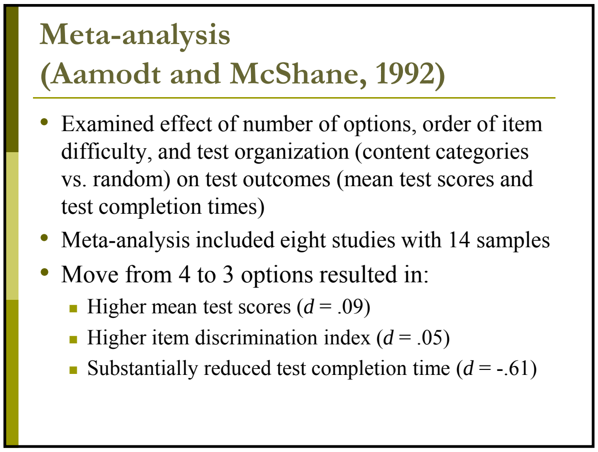
Sidick, Barrett, & Doverspike (1994)

- One of few published studies investigating number of MC test options in applied selection setting
- Compared 5-option test to 3-option test
  - Data obtained from two test administrations for entry-level police officers at a large municipality
- 3-option test created by eliminating options with lowest response rates from 5-option test
  - 5-option test administered 1991 (N = 1524)
  - 3-option test administered 1992 (N = 1790)



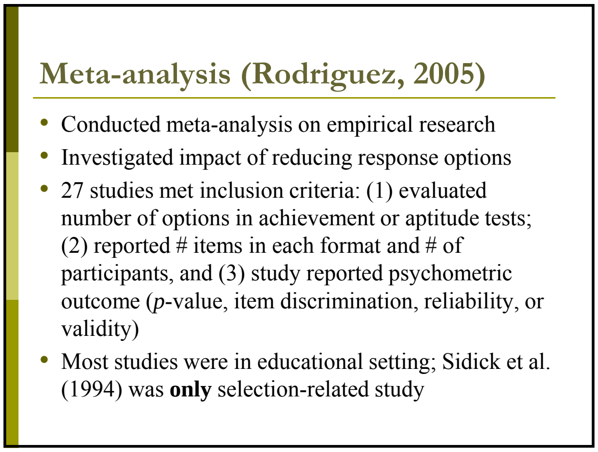
Sidick, Barrett, & Doverspike (1994)

- Reported higher reliability (α) for 3-option format (two of three subtests examined)
- Mean test scores were similar regardless of number of options
- One subtest had a *higher* mean score for 5-option format than for 3-option format
- “The overall results for the three-alternative test were similar to those obtained in previous test administrations in the same agency” (p. 833)



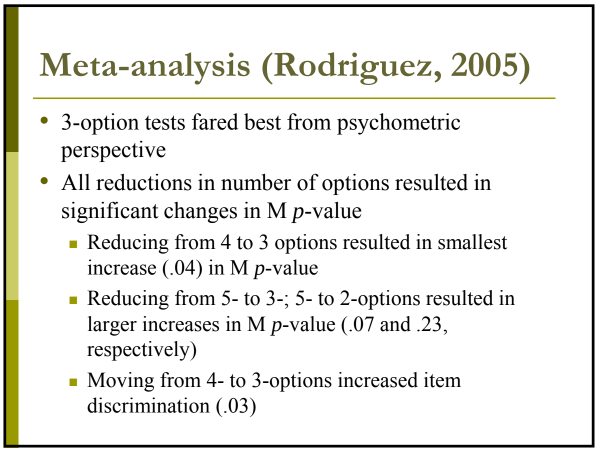
Meta-analysis (Aamodt and McShane, 1992)

- Examined effect of number of options, order of item difficulty, and test organization (content categories vs. random) on test outcomes (mean test scores and test completion times)
- Meta-analysis included eight studies with 14 samples
- Move from 4 to 3 options resulted in:
  - Higher mean test scores (d = .09)
  - Higher item discrimination index (d = .05)
  - Substantially reduced test completion time (d = -.61)



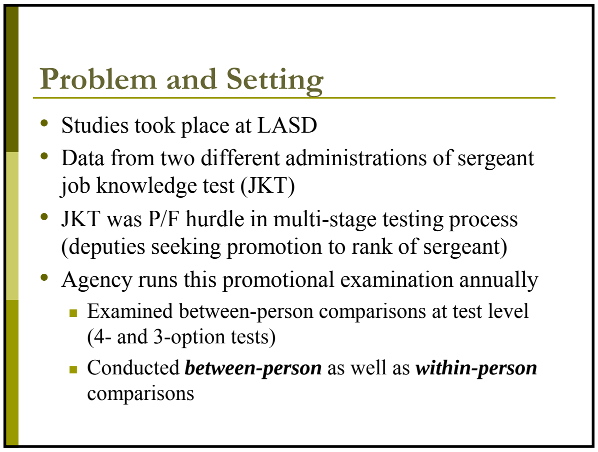
Meta-analysis (Rodriguez, 2005)

- Conducted meta-analysis on empirical research
- Investigated impact of reducing response options
- 27 studies met inclusion criteria: (1) evaluated number of options in achievement or aptitude tests; (2) reported # items in each format and # of participants, and (3) study reported psychometric outcome (*p*-value, item discrimination, reliability, or validity)
- Most studies were in educational setting; Sidick et al. (1994) was **only** selection-related study



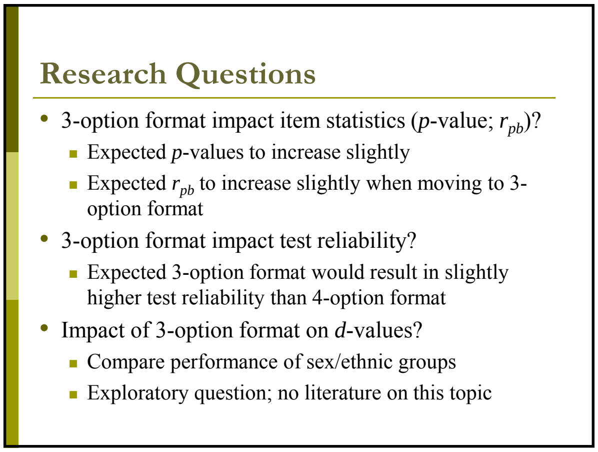
Meta-analysis (Rodriguez, 2005)

- 3-option tests fared best from psychometric perspective
- All reductions in number of options resulted in significant changes in M *p*-value
  - Reducing from 4 to 3 options resulted in smallest increase (.04) in M *p*-value
  - Reducing from 5- to 3-; 5- to 2-options resulted in larger increases in M *p*-value (.07 and .23, respectively)
  - Moving from 4- to 3-options increased item discrimination (.03)



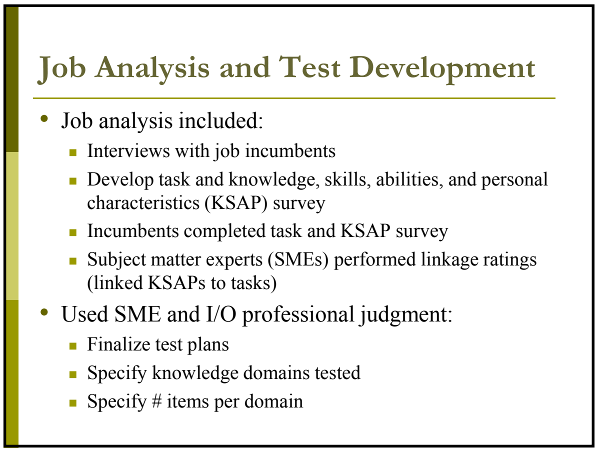
Problem and Setting

- Studies took place at LASD
- Data from two different administrations of sergeant job knowledge test (JKT)
- JKT was P/F hurdle in multi-stage testing process (deputies seeking promotion to rank of sergeant)
- Agency runs this promotional examination annually
  - Examined between-person comparisons at test level (4- and 3-option tests)
  - Conducted ***between-person*** as well as ***within-person*** comparisons



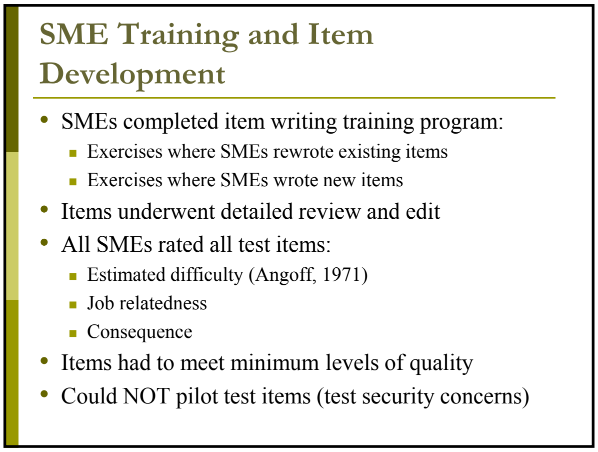
Research Questions

- 3-option format impact item statistics (*p*-value; r_{pb})?
  - Expected *p*-values to increase slightly
  - Expected r_{pb} to increase slightly when moving to 3-option format
- 3-option format impact test reliability?
  - Expected 3-option format would result in slightly higher test reliability than 4-option format
- Impact of 3-option format on *d*-values?
  - Compare performance of sex/ethnic groups
  - Exploratory question; no literature on this topic



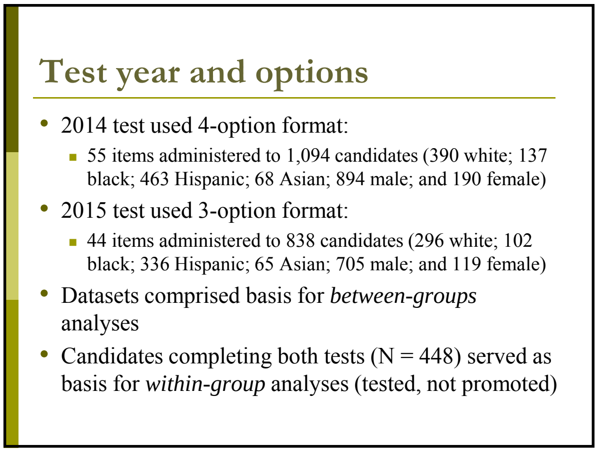
Job Analysis and Test Development

- Job analysis included:
  - Interviews with job incumbents
  - Develop task and knowledge, skills, abilities, and personal characteristics (KSAP) survey
  - Incumbents completed task and KSAP survey
  - Subject matter experts (SMEs) performed linkage ratings (linked KSAPs to tasks)
- Used SME and I/O professional judgment:
  - Finalize test plans
  - Specify knowledge domains tested
  - Specify # items per domain



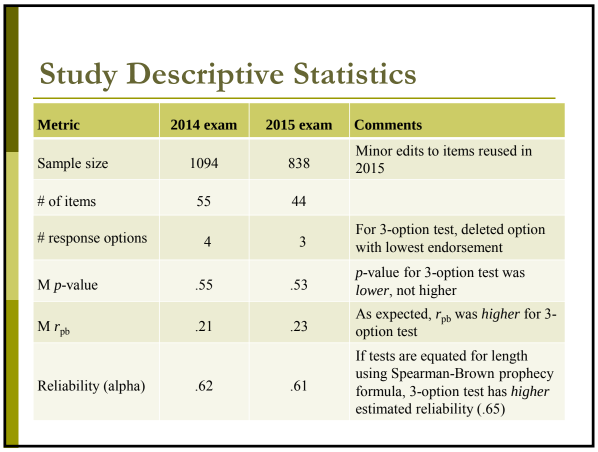
SME Training and Item Development

- SMEs completed item writing training program:
  - Exercises where SMEs rewrote existing items
  - Exercises where SMEs wrote new items
- Items underwent detailed review and edit
- All SMEs rated all test items:
  - Estimated difficulty (Angoff, 1971)
  - Job relatedness
  - Consequence
- Items had to meet minimum levels of quality
- Could NOT pilot test items (test security concerns)



Test year and options

- 2014 test used 4-option format:
  - 55 items administered to 1,094 candidates (390 white; 137 black; 463 Hispanic; 68 Asian; 894 male; and 190 female)
- 2015 test used 3-option format:
  - 44 items administered to 838 candidates (296 white; 102 black; 336 Hispanic; 65 Asian; 705 male; and 119 female)
- Datasets comprised basis for *between-groups* analyses
- Candidates completing both tests (N = 448) served as basis for *within-group* analyses (tested, not promoted)



Study Descriptive Statistics

Metric	2014 exam	2015 exam	Comments
Sample size	1094	838	Minor edits to items reused in 2015
# of items	55	44	
# response options	4	3	For 3-option test, deleted option with lowest endorsement
M *p*-value	.55	.53	*p*-value for 3-option test was *lower*, not higher
M r_{pb}	.21	.23	As expected, r_{pb} was *higher* for 3-option test
Reliability (alpha)	.62	.61	If tests are equated for length using Spearman-Brown prophecy formula, 3-option test has *higher* estimated reliability (.65)

Between groups (second to last row) & Within-group (bottom row) *d*-values

2014 JKT (4-option)				2015 JKT (3-option)				Difference (2015 – 2014)			
W-B	W-H	W-A	M-F	W-B	W-H	W-A	M-F	W-B	W-H	W-A	M-F
0.14	0.19	-0.06	0.28	0.41	0.42	0.20	0.40	0.27	0.23	0.26	0.12
0.25	0.23	0.16	0.36	0.28	0.35	0.15	0.50	0.03	0.12	-0.01	0.14

Within-group analysis

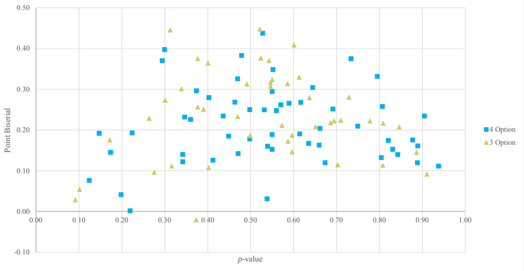
- Dependent *t*-test (compared *M* score by year)
  - Result was *ns* ($t = -0.399$; $df = 445$; $p = 0.690$)
- Scores on two tests were moderately correlated ($r = .47$; $p < .001$ two-tailed)
- Because people who retested did not promote, range restriction might be issue
  - After correction (direct restriction), *r* increased to .52
  - Range restriction was *not* a major issue in reducing this correlation

Comparing *p*-value and r_{pb}

- Figure 1 [next slide] compares *p*-value and r_{pb} for items in both tests
- At low *p*-values (.10 to .20), and high *p*-values (.80 to .90), r_{pb} tends to be low
- At moderate *p*-values (.30 to .70), r_{pb} tends to be higher
  - Curvilinear relationship
  - Expected outcome because with dichotomous item (0/1), variance is maximized at $p = q = .50$
  - With greater variance, expect higher correlation (r_{pb})

p-value and r_{pb}

Figure 1. Relationship of *p-value* and r_{pb} for 4- and 3-option items

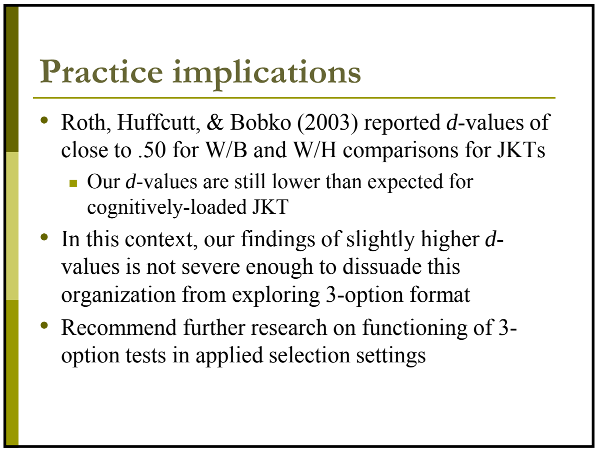


Research implications

- In our applied study, *M* test score did **not** increase when moving from 4- to 3-option test
  - Found *lower* mean test score on 3-option test (between Ss)
  - Found *lower* mean test score (within-Ss analysis)
- We did find increase in mean r_{pb} for 3-option test
- Why do results differ from other studies?
  - Most empirical studies based on academic tests
  - Selection tests may function differently
  - Planning further comparisons (sergeant, lieutenant, multiple civilian tests)

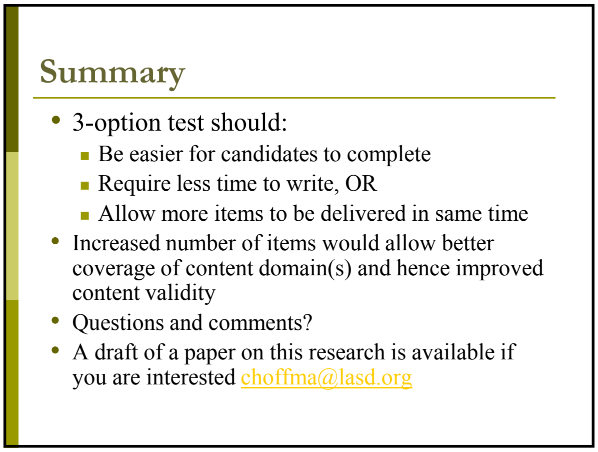
Practice implications

- Practical benefits from 3-option format
  - Reduced test development time
  - Reduced test administration time
- Finding that 3-option test had slightly higher sex- and ethnic-group *d*-values raises some concern
- I/O psychology has ongoing (and not very successful) goal of reducing adverse impact in valid predictors
  - Finding that intervention might contribute to larger mean group differences is not what one hopes for!



Practice implications

- Roth, Huffcutt, & Bobko (2003) reported *d*-values of close to .50 for W/B and W/H comparisons for JKTs
  - Our *d*-values are still lower than expected for cognitively-loaded JKT
- In this context, our findings of slightly higher *d*-values is not severe enough to dissuade this organization from exploring 3-option format
- Recommend further research on functioning of 3-option tests in applied selection settings



Summary

- 3-option test should:
  - Be easier for candidates to complete
  - Require less time to write, OR
  - Allow more items to be delivered in same time
- Increased number of items would allow better coverage of content domain(s) and hence improved content validity
- Questions and comments?
- A draft of a paper on this research is available if you are interested choffma@lasd.org