Category Ratings and Assessments:
Impact on validity, utility, and managerial choice

Jeffrey M. Cucina, Ph.D.
U.S. Customs and Border Protection
Henry Busciglio, Ph.D.
U.S. Customs and Border Protection
Kathlea Vaughn, M.A.
U.S. Customs and Border Protection

Presented at the Annual Conference of the International Personnel Assessment Council
Tuesday, July 19, 2011, Washington, DC.

U.S. Customs and Border Protection

Opinions expressed are those of the authors and do not represent the position of U.S. Customs and Border Protection

Today's Presentation

- Description of top-down selection and category ratings
- Description of data analyses and methodology
  - Used real and simulated data
- Presentation of five research questions with results
- Conclusions, recommendations, and topics for practitioners to consider

Federal Government Selection

- Applicants compete for positions based on their knowledge, skills, and abilities
- Traditionally, applicants are rank-ordered using assessment scores (from 70-100) and hiring is top-down
- Recent Presidential Memorandum (November 2010) included switch to category ratings
  - Can loosely be described as a form of banding

Purpose of Study

- Category ratings have become a hot topic among HR professionals, hiring managers, and media outlets covering Federal issues
- We could find no past published/presented research addressing category ratings
- Testing professionals in the Federal Government need to convert raw test scores into category ratings

Focus of Study

- Large-scale mission critical occupations.
  - Hundreds of openings, thousands of incumbents, tens of thousands of applicants
  - Often use a professionally developed and validated test battery
  - Federal agencies that hire assessment professionals usually have them to focus on these large occupations
  - Focus of this study
- Small occupations not examined in our study
  - One opening, 5-10 applicants

Top-Down 70-100 Explained

- Raw test scores are "transmuted" to 70-100 scale
  - Linear transformation
  - 70 is required passing/cutoff score. Failing applicants do not receive transmuted score.
  - Veterans can receive 5 or 10 bonus points
  - Hiring is top-down

Rule of Three Explained

- Hiring manager choose among top 3
  - Hiring manager can make an offer to any of the top 3 applicants (based on 70-100 scores)
  - If multiple offers are made, then new groups of 3 are created
    - Occurs if an applicant declines an offer or >1 position to be filled
  - If an applicant is passed over three times (i.e., appears in the top 3 but never made an offer) he or she is automatically eliminated
    - Unless he or she has veterans' preference

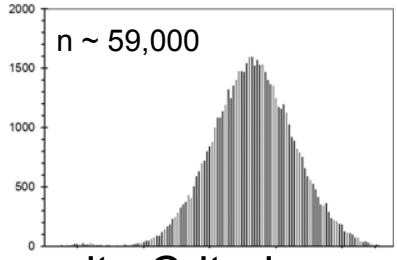
Category Ratings Explained

- Raw test scores are placed into categories
  - Three categories are the most common
    - Highly-Qualified (top)
    - Well-Qualified (middle)
    - Qualified (bottom)
  - Hiring manager can choose any applicant within a category (ignoring veterans' preference)
  - Proposed as an alternative to the rule of three as part of Federal hiring reform
  - Test scores used to place applicants into categories
  - Can merge categories when ≤2 applicants in one category

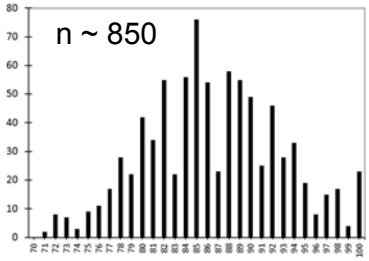
Method: Predictor Tests Used

- Composite Predictor (validity of .43)
  - Archival applicant data was used
  - For purposes of this study, we created a composite variable of a cognitive measure and non-cognitive measure

Raw Test Score



Transmuted 70-100 Score



- Composite Criterion
  - Training academy scores (also used as separate criterion)
  - Task-Based Job Simulation Scores
  - Supervisory Ratings

Method: Creating Category Ratings

- Officially, must use a job analysis (more on this later)
- We used six different approaches

- Best Case Scenario Categories
  - Used an empirical method, which maximizes criterion-related validity
  - Two cut scores used were those with the highest r_{pbi} with job performance
  - These cut scores resulted in the following predictor score ranges for each Category Rating

Rating	Score Range
3	91 - 100
2	84 - 90
1	70 - 83

Method: Creating Category Ratings

- Decades Categories – Based on transmuted scores
  - Category 1 = 70s
  - Category 2 = 80s
  - Category 3 = 90s-100
- Tertiles – Top, Middle, and Bottom Thirds
  - Similar to quartiles or quintiles, but with three groups

Method: Creating Category Ratings

- Worst Case – Negative Skew
  - Based on transmuted scores
  - Category 1 = 70
  - Category 2 = 71
  - Category 3 = 72-100
- Worst Case – Middle
  - Based on transmuted scores
  - Category 1 = 70
  - Category 2 = 71 through 99
  - Category 3 = 100
- Worst Case – Positive Skew
  - Based on transmuted scores
  - Category 1 = 70-98
  - Category 2 = 99
  - Category 3 = 100

Method: Datasets

Large Applicant Dataset

- n ~ 59,000
- Represented all applicants taking one particular form/series
- Raw test scores normally distributed

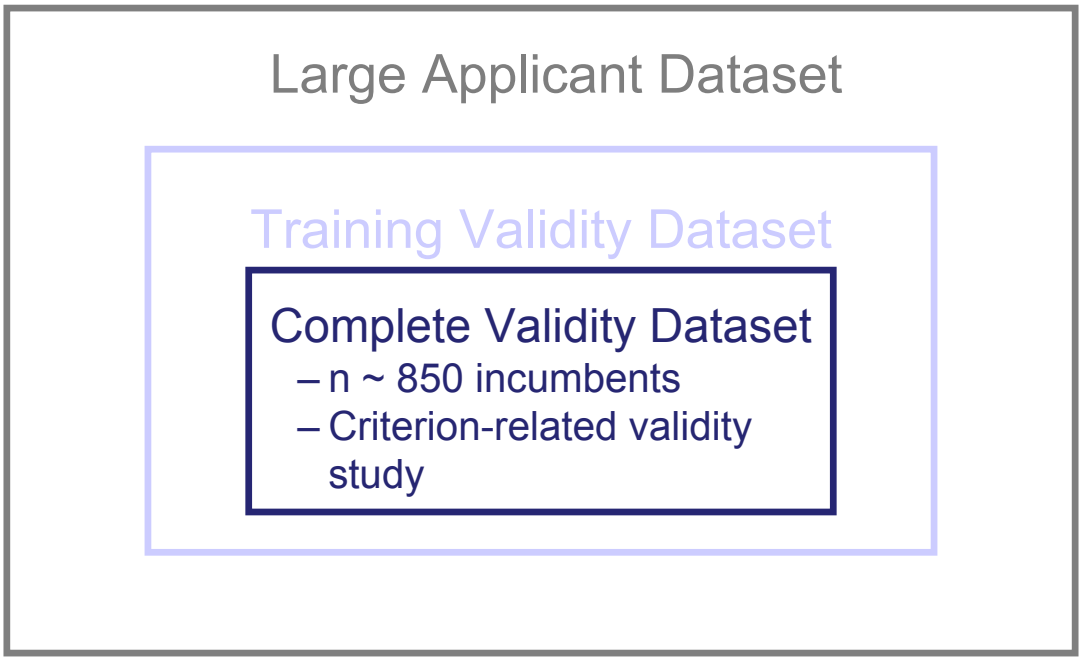
Method: Datasets

Large Applicant Dataset

Training Validity Dataset

- Subset of Large Applicant Dataset
- n ~ 6,000
- Applicants who were hired and went to training academy
- Training performance criterion

Method: Datasets



1. What is the impact of category ratings (vs. top-down 70-100 rankings) on criterion-related validity?
   - MacLane (2010) hypothesized decrease in validity
   - We concur and hypothesize that validity will decrease
   - Used complete validity dataset
   - Correlated composite criterion with transmuted 70-100 scores, and category ratings

1. Category Ratings → Lower Validity

Predictor/Method	$r_{Uncorrected}$	*p*
Raw Test score	**.430**	**< .001**
Transmuted 70-100	**.429**	**< .001**
Categories		
- Best Case	**.414**	**< .001**
- Decades	**.374**	**< .001**
- Tertiles	**.335**	**< .001**
- Worst Case Positive Skew	**.164**	**< .001**
- Worst Case Middle	**.158**	**< .001**
- Worst Case Negative Skew	.053	.125

1. Category Ratings ignore valid information

- Within each category, the transmuted score was statistically significant. (note: †: p = .056)

Predictor/Method	Validity of Transmuted Score within:		
	Category 1 (Bottom)	*Category 2 (Middle)*	*Category 3 (Top)*
Categories			
- Best Case	**.162****	**.117***	**.166***
- Decades	**.185†**	**.230****	**.232****
- Tertiles	.081	.089	**.332****
- Worst Case Positive Skew	**.407****	(constant)	(constant)
- Worst Case Middle	(constant)	**.409****	(constant)
- Worst Case Negative Skew	(constant)	(constant)	**.427****

Note: †: p = .056.

1. Category Ratings → Decremental Validity

- Conducted hierarchical linear regression; Step 1: Category Rating Score; Step 2: Transmuted Score
- Transmuted score **always** added incremental validity
- Using category ratings instead of transmuted has decremental validity

Predictor/Method	ΔR^2 vs. Transmuted	*p*
Categories		
- Best Case	**- .018**	**< .001**
- Decades	**- .045**	**< .001**
- Tertiles	**- .072**	**< .001**
- Worst Case Positive Skew	**- .157**	**< .001**
- Worst Case Middle	**- .159**	**< .001**
- Worst Case Negative Skew	**- .181**	**< .001**

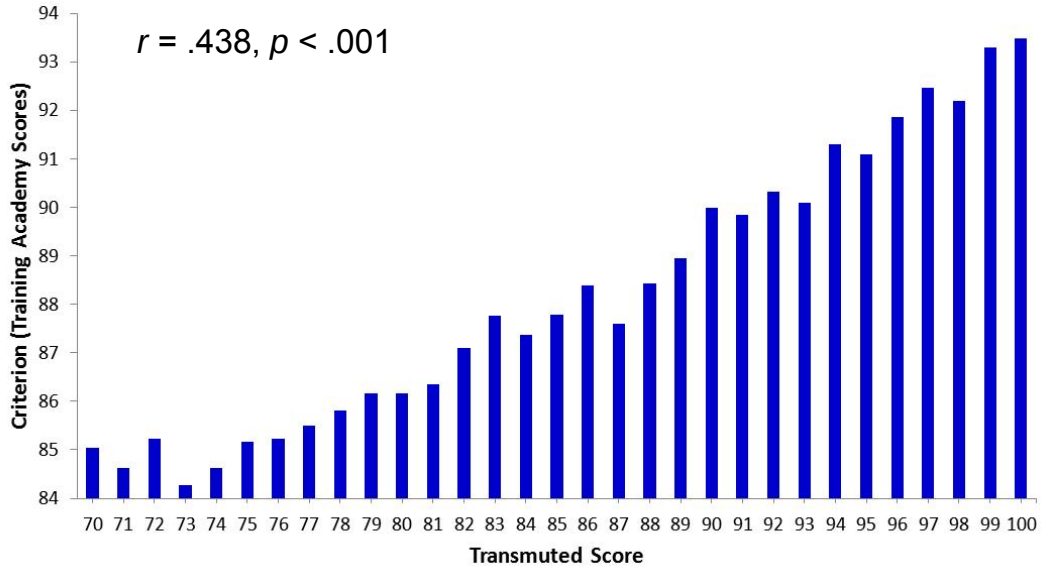
1. Conclusion

- Category ratings do decrease validity
- Amount of decrease in validity depends on how categories are formed
- Consistent with MacLane's (2010) hypothesis

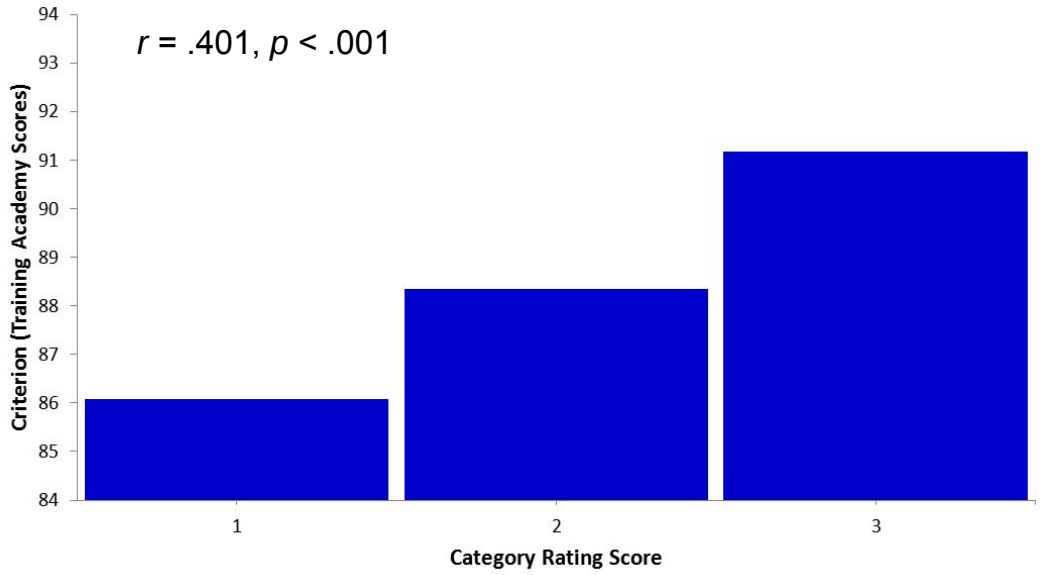
2. What is the impact of category ratings on merit; in other words, are the top applicants (in terms of criterion scores) always selected?

- Two hypotheses (drawn from banding literature)
  - Pro-Banding Hypothesis – Differences in transmuted scores within a category are largely due to chance and not meaningful
  - Anti-Banding Hypothesis – Differences in transmuted scores are meaningful, especially with large pools of applicants
    - See OPM white paper by Frank Schmidt (no date)
- Used training validity dataset

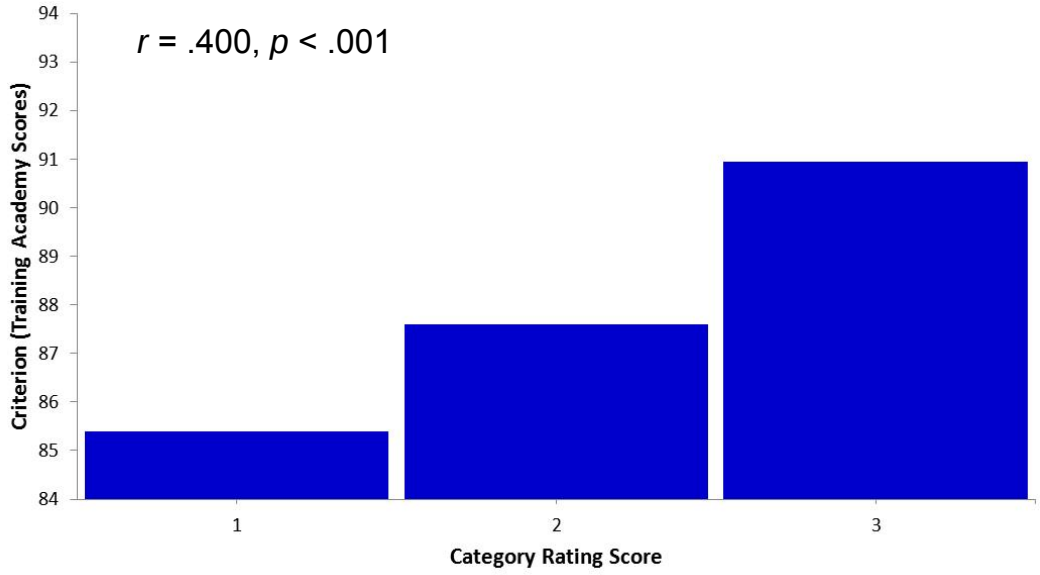
2. Average criterion score for applicants at each transmuted score



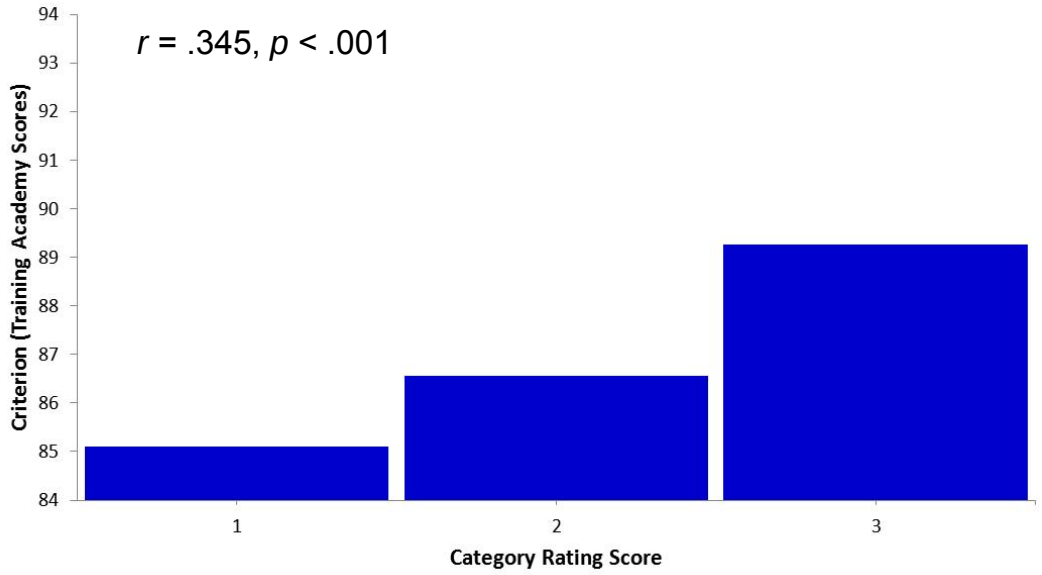
2. Average criterion score for applicants at each category: Best Case Categories



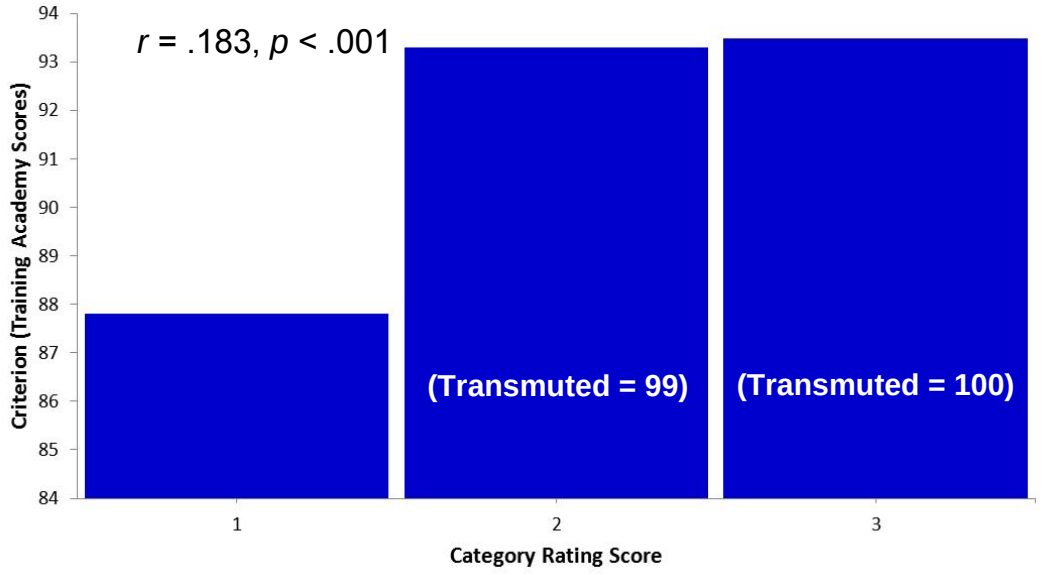
2. Average criterion score for applicants at each category: Decades Categories



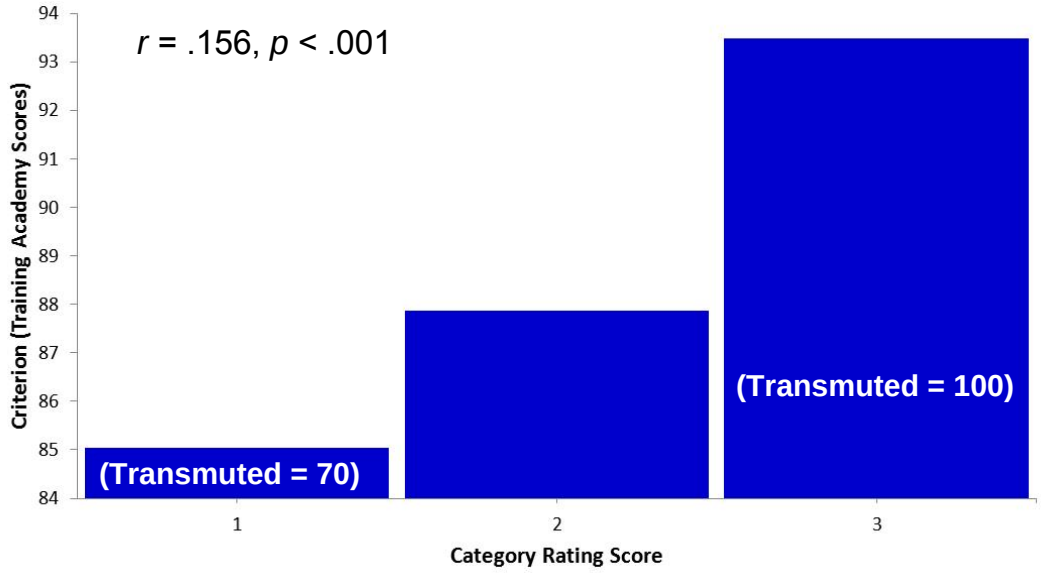
2. Average criterion score for applicants at each category: Tertiles Categories



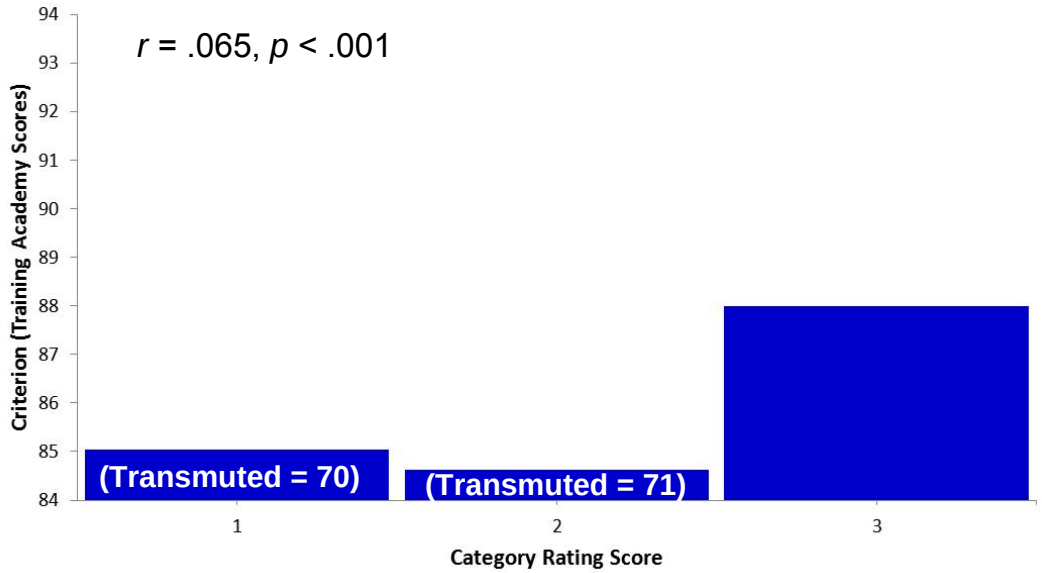
2. Average criterion score for applicants at each category: Worst Case Positive Skew Categories



2. Average criterion score for applicants at each category: Worst Case Middle Categories



2. Average criterion score for applicants at each category: Worst Case Negative Skew Categories



2. Conclusion

– Using transmuted score allows for finer distinctions among applicants on the criterion

– Using category ratings erases the finer distinctions → Applicants with (slightly) lower criterion scores may be selected ahead of those with (slightly) higher criterion scores

3. What is the impact of category ratings on utility, compared to using transmuted scores?

- Key benefit of testing is return on investment via better quality hires
- Compared change in utility when moving from transmuted scores to category ratings
- Used below utility formula and assumptions

$$\Delta U = T \bullet N_s \bullet (r_1 - r_2) \bullet SD_Y \bullet z \; - \; \frac{N_s \bullet (C_1 - C_2)}{p}$$

T = Tenure in years of average selectee = 20 years (agent hired by age 37 retires at age 57 = 20 years)

N_s= Number selected per year = 1,000 (same as congressionally mandated FY11 hiring goal for Border Patrol)

r_1 = Validity of new selection system (e.g., category ratings)

r_2 = Validity of old selection system (e.g., transmuted)

$SD_Y \bullet z$ -= Dollar value of performance = .32 (medium complexity job) • $60,274 (GS-12-Step-1)

z = mean score of those who were selected = 0.78… (used for both transmuted and category ratings)

C_1 = Cost of old selection system = C_2 = cost of new selection system = N/A (cancels out)

p = selection ratio = N/A (cancels out)

3. Category Ratings → Lower Utility

Predictor/Method	*Change in Dollars*
Transmuted (vs. raw score)	-$301,034
Categories (vs. transmuted)	
- Best Case	**-$4,515,522**
- Decades	**-$16,556,915**
- Tertiles	**-$28,297,273**
- Worst Case Positive Skew	**-$79,774,228**
- Worst Case Middle	**-$81,580,437**
- Worst Case Negative Skew	**-$113,189,094**

- Conclusion: Category ratings reduces return on investment

4. What is the impact of category ratings on veterans' preference?

– Refresher on veterans' preference

TP Veterans - Preference eligibles with no disability rating
- Receive 5 points under rule of three

XP Veterans - Disability rating less than 10%
- Receive 10 points under rule of three

CP Veterans - Disability rating of at least 10% but less than 30%
- Receive 10 points and move to very top of list

CPS Veterans - Disability rating of 30% or more
- Receive 10 points and move to very top of list

4. What is the impact of category ratings on veterans' preference?

- Rule of Three
  - Veterans receive an extra 5 (TP) or 10 (XP) points that is added to their raw 70-100 transmuted score
    - Yields scores ranging from 70 to 110 (for all applicants)
    - If there are ties, then veterans listed first
- Category Ratings
  - Within a category, TP (5-point) and XP (10-point) veterans now move to the top of their original category and must be hired first (if hiring made from that category)
  - CP and CPS move out of their category (if necessary) to the top of the top category

4. Two TP (5-point) veterans under Decades model

- Rule of Three

Veteran A: 90 $\xrightarrow[\text{Pref.}]{\text{Add Vets.}}$ Veteran A: 95

Moves ahead of non-veterans with scores of 90-95

Veteran B: 89 $\xrightarrow[\text{Pref.}]{\text{Add Vets.}}$ Veteran B: 94

Moves ahead of non-veterans with scores of 89-94

4. Two TP (5-point) veterans under Decades model

– Category Ratings

Vet. A: 90 → Assign to Category → Highest Category → Add Vets. Pref. → Top of Highest Category

Moves ahead of non-veterans with scores of 90-100

Vet. B: 89 → Assign to Category → Middle Category → Add Vets. Pref. → Top of Middle Category

Now only moves ahead of non-veterans with scores of 89. Unlike rule of three, now behind 90-94.

4. Practical Significance: How many applicants would really be impacted by this?

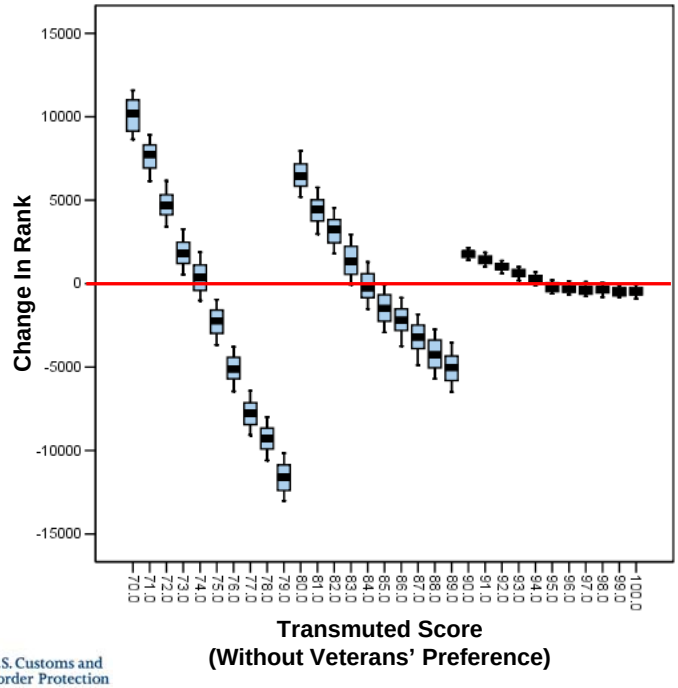
– Used large applicant dataset

– Rank-ordered applicants under decades category ratings model vs. 70-100

– Added veterans' preference points and moved floaters to top

– Used a random number to rank-order applicants with ties (same random number used for both scenarios)

4. Average change in rank (category ratings vs. rule of three) for TP (5-point) veterans by transmuted score

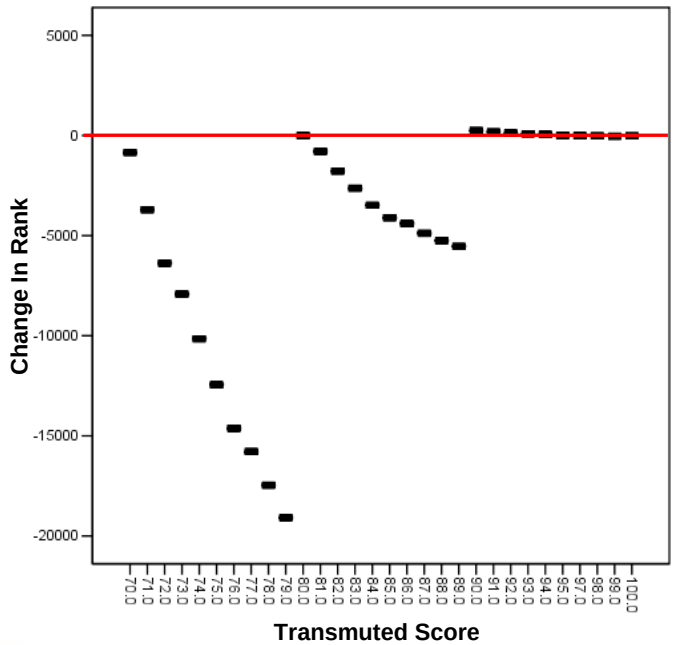


- Applicants above red line were ranked higher under category ratings
- Applicants below red line were ranked lower under category ratings

Note: To create this chart, we split the datafile by transmuted score and computed the average change in ranking (i.e., rule of three rank – category ratings rank) for veterans with each raw score under the decades model.

U.S. Customs and Border Protection

4. Average change in rank (category ratings vs. rule of three) for XP (10-point) veterans by transmuted score



- Applicants above red line were ranked higher under category ratings
- Applicants below red line were ranked lower under category ratings

U.S. Customs and Border Protection

4. Results: TP (5-point) veterans

Under category ratings (vs. rule of three):

Veterans ranked higher: 3,483 (48%)

Veterans ranked lower: 3,756 (52%)

Veterans ranked same: 0 (0%)

Average change in rank: -637 places

Range of change in rank

- Largest drop: -13,011 places
- Largest gain: 11,589 places

Wilcoxon signed-rank test: $Z = -7.706$; $p = .001$

4. Results: XP (10-point) veterans

Under category ratings (vs. rule of three):

Veterans ranked higher: 42 (16%)

Veterans ranked lower: 216 (84%)

Veterans ranked same: 0 (0%)

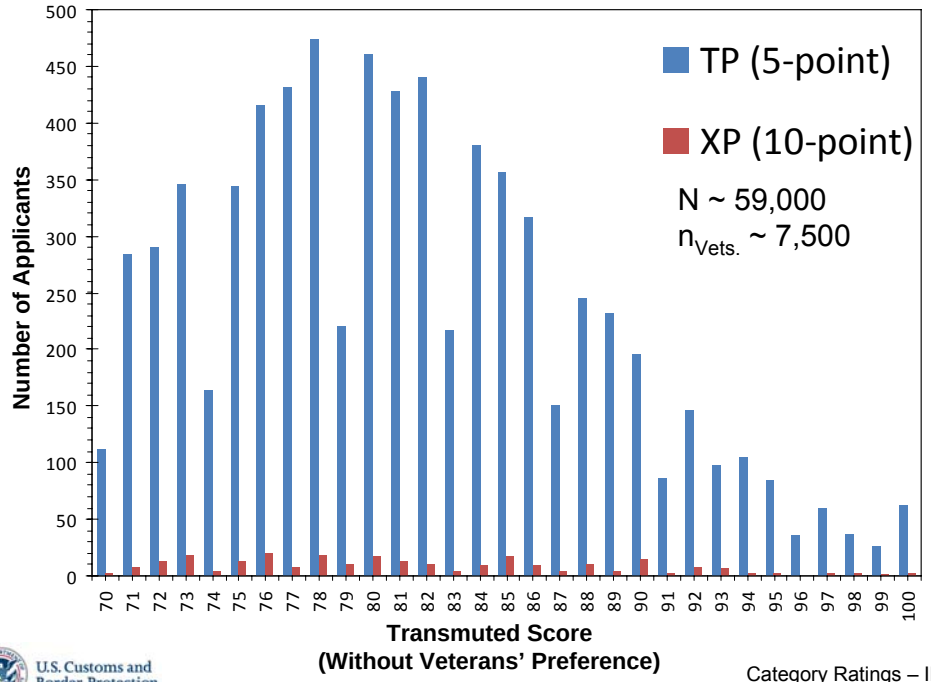
Average change in rank: -6,499 places

Range of change in rank

- Largest drop: -19,135 places
- Largest gain: 250 places

Wilcoxon signed-rank test: $Z = -12.642$; $p < .001$

4. Number of TP (5-point) and XP (10-point) veterans by transmuted score



4. Why this could matter... (Veterans' preference is popular topic in the courts)

- Consider recent court cases over veterans' preference
  - The Federal Career Intern Program (FCIP) was recently struck-down as written by an Administrative Law Judge (ALJ) at the Merit Systems Protection Board (MSPB) (*Dean v. OPM and Evans v. Department of Veterans Affairs, 2010, MSPB 213*)
    - FCIP didn't require a public job posting
    - An agency used FCIP to circumvent hiring a veteran
    - ALJ ruled that this could prevent veterans from being hired and was not legal
  - OPM's ALJ exam had a legal challenge involving score compression (0-100 vs. 70-100) and 5 vs. 10-point preference. (*Azdell and Fishman v. OPM, 2003, SCOTUS 03-624*)

4. Conclusion

- Category ratings changes the nature of veterans' preference
  - Some veterans do better, but others do worse
- Some veterans who would be hired under rule of three but not under category ratings
- Which veterans get ranked higher and which do not is somewhat arbitrary
  - Is this in the spirit of the law?
  - Is this fair?
  - (These are points to ponder)
  - (Note, none of us have a J.D.)

5. What is the impact of category ratings on managerial choice?

(Moving away from veterans and validity to new a topic...)

–Often cited benefit of category ratings is that hiring manager can choose anyone within a category (ignoring veterans' preference)

–Categories can be combined when 2 or fewer applicants remain in the higher category

- If higher category did not have 2 applicants at first, then all but 2 must have been offered a position.
- Applicants that hiring manager didn't choose are still counted

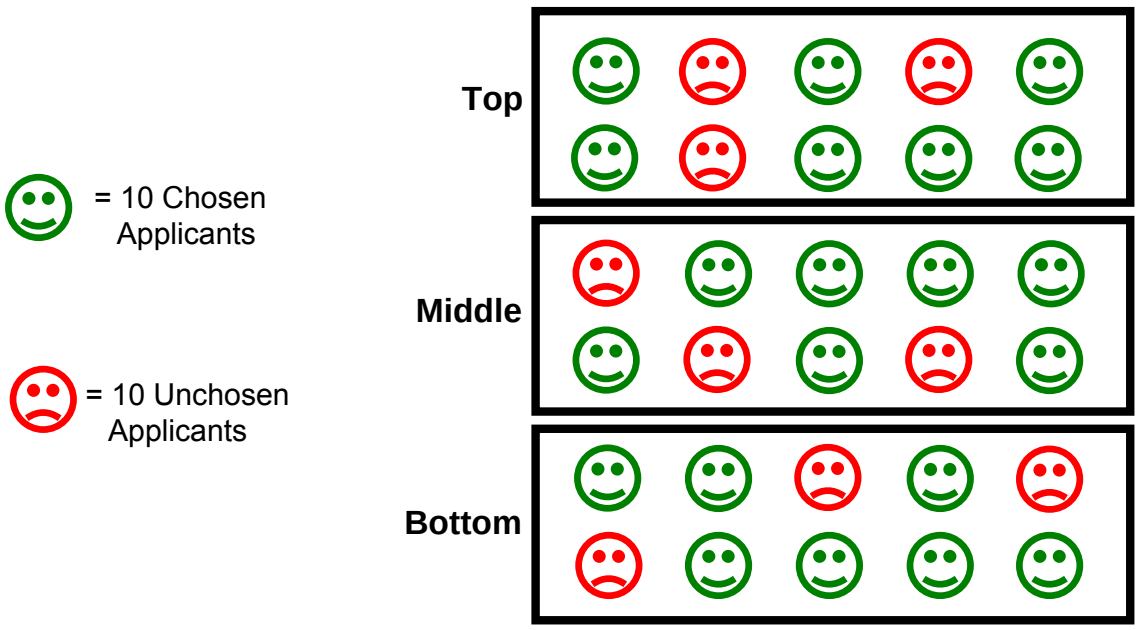
–With large occupations, will need to fill more positions than candidates in highest category

- We propose that the rule of three may lead to better managerial choice in these situations

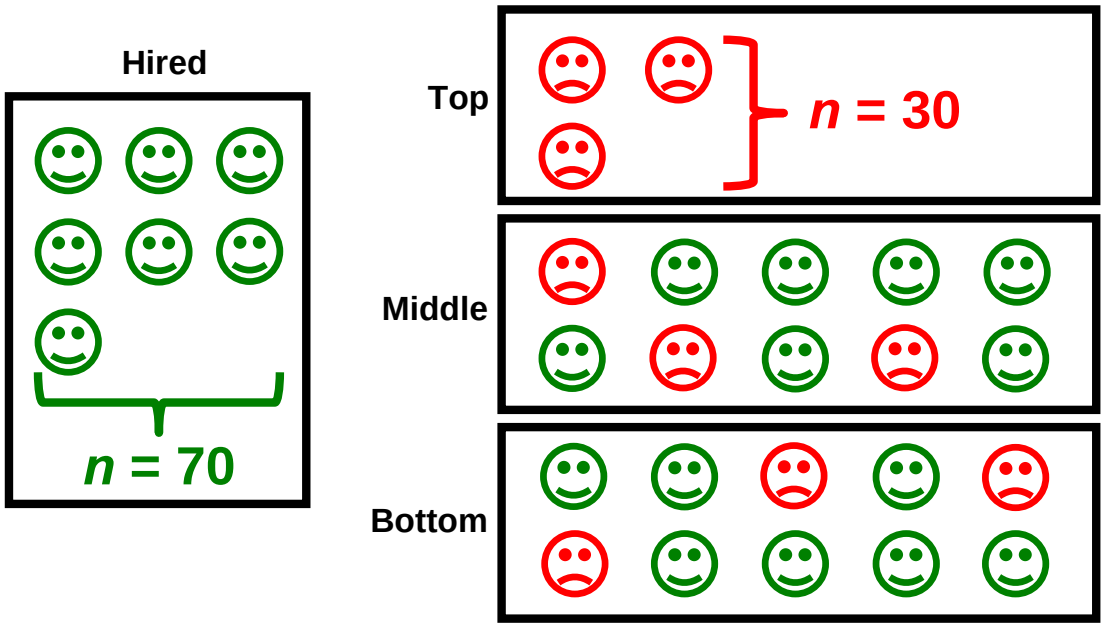
5. Scenario to consider

- Imagine applicants assessed using a measure that either has lower validity or misses important competency for the job
- There are 300 applicants, in three categories of 100 applicants each
- Hiring Manager does not want to hire 30% of the applicants (for whatever valid or invalid reason)
  - (In each category, 30 of the 100 applicants are unchosen by hiring manager)
- Hiring goal is to hire 150 applicants

5. A graphical depiction



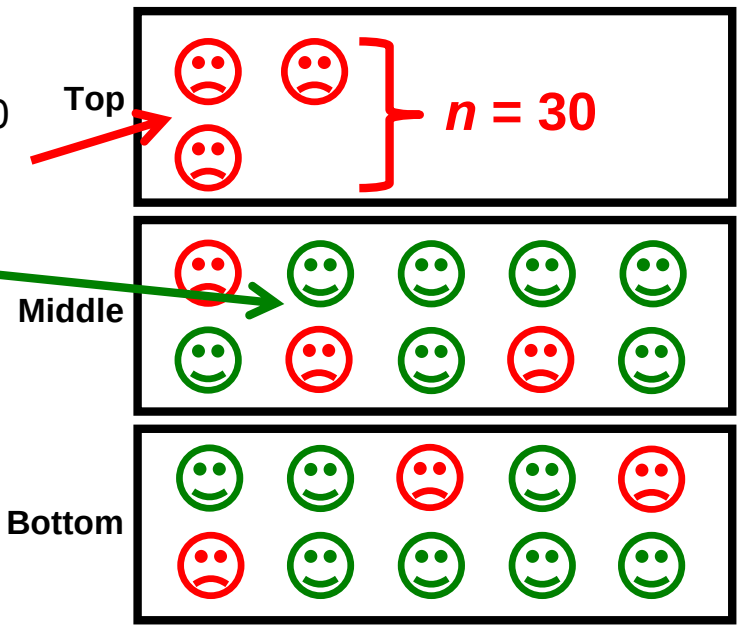
5. Make all top category job offers...



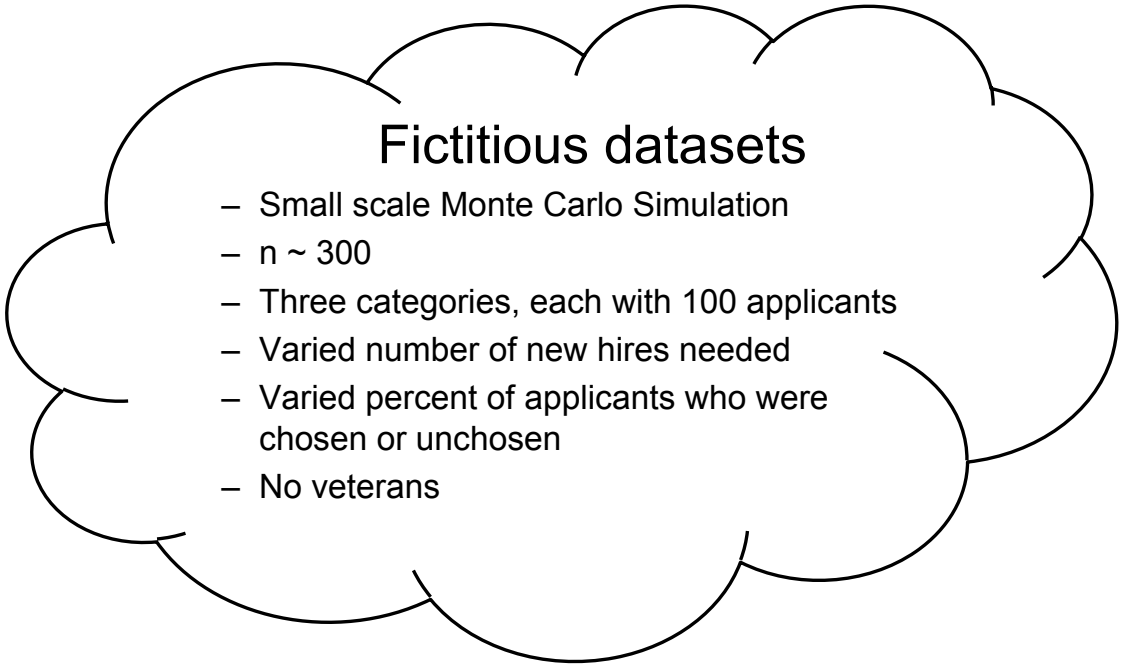
5. We still need to hire 80 more applicants...

Must hire 28 of the 30 unchosens from top, before hiring from middle

Plan B: Start from scratch with a new announcement. (Hope that 30 top unchosens don't reapply.)



Method: Datasets



5. Category Ratings vs. Rule of Three: Percent of Unchosen Applicants Discarded

- Similar situation with 3 categories of 100 applicants each
- In the table below we vary the percent of unchosen applicants

Hiring Goal: Only from Top Category

Unchosen	Category Ratings	Rule Of Three
10%	**100%**	**100%**
20%	**100%**	92%
30%	**100%**	83%
40%	**100%**	74%
50%	**100%**	48%

Unchosen	Category Ratings	Rule Of Three
60%	**100%**	56%
70%	**100%**	33%
80%	**100%**	24%
90%	**100%**	9%

Columns 2 & 3 show percentage of unchosen applicants not selected (i.e., able to be passed over)

5. Category Ratings vs. Rule of Three: Percent of Unchosen Applicants Discarded

Hiring Goal: 150

Unchosen	Category Ratings	Rule Of Three
10%	77%	**100%**
20%	63%	**93%**
30%	66%	**83%**
40%	71%	**74%**
50%	**64%**	56%

Unchosen	Category Ratings	Rule Of Three
60%	**64%**	45%
70%	**57%**	33%
80%	**56%**	25%
90%	**52%**	11%

5. Category Ratings vs. Rule of Three: Percent of Unchosen Applicants Discarded

Hiring Goal: Everyone (but unchosens)

Unchosen	Category Ratings	Rule Of Three
10%	63%	**100%**
20%	68%	**92%**
30%	58%	**77%**
40%	58%	**68%**
50%	**61%**	52%

Unchosen	Category Ratings	Rule Of Three
60%	**63%**	44%
70%	**66%**	32%
80%	**64%**	22%
90%	**66%**	10%

5. Conclusion

- Category ratings approach maximizes managerial choice when selections are limited to candidates in the top category
- Rule of three approach maximizes managerial choice when categories are collapsed
  - Except when 50% or more of candidates are unchosen, then category ratings approach maximizes managerial choice
- Rule of three approach may give more managerial choice for large occupations with mass hiring
  - Since categories must be collapsed to meet hiring goals
- Category ratings could give more managerial choice for small occupations with few hires
  - Since hiring will take place only from top category

Things to Think About

- Cutoff scores for categories
  - Must be created before job is posted
  - Must be created using job analysis
    - Per OPM regulations and *Delegated Examining Operations Handbook*
  - How to set legally defensible cutoff scores on an objectively scored multiple-choice test?
    - Traditionally created using criterion-related validation study, Angoff standard setting study, etc.
      - This is not a "job analysis" as described in the literature
    - Some job analysis surveys include rating scales that parallel benchmarks for competency-based rating scales used in structured interview, KSA-essay panel review, etc.
      - Linking this job analysis survey data to multiple-choice test scores would require validation or standard setting study

Things to Think About

- Might be good to read *Lewis v. Chicago, 2011*
  - Case really involved category ratings
  - Reached Supreme Court over a time-to-file issue
  - Remanded to Seventh Circuit, which decided for the plaintiffs on May 13, 2011
    - Decision mentioned choice of 89 as cutoff was "not justified" and method "did not follow the common civil-service practice of hiring in rank order from a list"
    - Caveats:
      - Seventh circuit, not Federal circuit (which covers Federal hiring)
      - Could be appealed
      - Should check with your agency's counsel

Things to Think About

- Model category ratings policies only provide hiring managers with names of candidates in a category
  - Template policies do not allow hiring officials to view test scores or other information
  - Hiring managers may receive names (and nothing more) of 100s or 1000s of applicants
  - Could interview any applicant, but to interview all could be laborious

Things to Think About

- Model category ratings policies only provide hiring managers with names of candidates in a category (cont'd.)
  - Research from the resume literature has shown that names can introduce disparate treatment toward minority groups
    - Field experiment found that fictitious resumes with "White-sounding names" received 50% more call-backs than those with "African-American sounding names," despite identical content
      - Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*(4), 991-1013.

Things to Think About

- Hiring managers told they can select "anyone"
  - What about *Merit Principles*?
    - "selection and advancement should be determined solely on the basis of relative ability, knowledge and skills" (5 USC 2301)?
  - What about *Prohibited Personnel Practices*?
    - "political affiliation, race, color, religion, national origin, sex, marital status, age, or handicapping condition" (5 USC 2301)
    - "nepotism" (5 USC 2302)
    - "factors other than personal knowledge or records of job-related abilities or characteristics" (5 USC 2302)

Other Ideas

- ✗ Use 31 categories to match 70-100 scale
  - Would nearly eliminate veterans' preference
- ✓ Give HR staff, Personnel Research Psychologists, and Hiring Managers choice between rule of three and category ratings?
- ✓ Introduce a rule of 5, 7, or 10 instead of 3
- ✓ Provide test scores to selecting officials
- ✓ Request exemption from OPM (see President's Memorandum Section 5 (d)

Audience Ideas?

Questions and Comments from the Audience