Untangling the Effects of Candidate and Assessor Race and Gender on Assessment Center Ratings:
A Story, An Important Issue,
and A Research Study

Alyssa M. Gibbons

Adam J. Vanhove

George C. Thornton III

Colorado State University

The Problem

- Increasing diversity in many areas of the world leads to concerns about discrimination and fairness.

  - Race/ethnicity, Gender, Religion, Etc.

- Avoiding discrimination in employment decisions is legally mandated in some nations.

  - Motivated by moral, social, and political concerns in others.

ACs: Part of the Solution?

- ACs are often touted as a fairer alternative to other kinds of selection procedures.
  - Courts often recommend ACs over other methods.
- Research on demographic effects in ACs has often found:
  - No differences
  - Small differences
- Where differences are present, they are often smaller than those from other selection methods.

ACs: Part of the Solution?

- ACs show less adverse impact (systematic group differences) than many other selection methods.

- ACs are perceived as fair by candidates.
  - Clearly job relevant.
  - Opportunity to display one's skills and abilities.

The Not-So-Good News

- *Less* adverse impact ≠ *no* adverse impact.
- Recent research does show racial and gender differences in some AC ratings.
  - Cross-cultural ACs may face even more issues.
- Variability among ACs.
  - Some show differences, some don't.
- Can't assume that every AC is perfectly fair.
- Perceptions of fairness may be just as important as actual fairness.

Potential for Bias in ACs

- Assessors: human beings making judgments about others.
  - Human judgment introduces potential for personal bias
- Do assessors bring their biases with them?
- Can natural biases be controlled?
- Do candidates believe that assessors are biased?

Fighting Bias

- AC practices combat bias in many ways:
  - Multiple observation opportunities
  - Multiple assessors
  - Rotation of assessors
  - Assessor training
  - Focus on behavior
  - Clearly defined rating standards
  - Integration and consensus processes
- Do they work? Are these practices enough?

Three Kinds of Potential Bias

- Main effects for assessees.
  - Do members of the majority group get higher ratings?
  - Previous research: sometimes, but effects are small.
- Main effects for assessors.
  - Do men/women, older/younger, engineers/sales give higher ratings?
- Interaction between candidate and assessor characteristics.
  - Do assessors favor candidates who are similar to themselves?
  - Previous research is lacking in AC contexts.

Similarity Bias

- In performance appraisal and other contexts, raters **do** seem to favor those who are demographically similar.
  - Practical effect of these biases may be small.
- Reasonable to expect similar findings in ACs.
  - But effects may be larger due to the relatively short duration of an AC.
- Implications for AC practice:
  - Assessor training.
  - Assignment of candidates to assessors.

Questions

- Does demographic similarity between assessor and candidate affect AC ratings?

- If so, are these effects enough to change candidate outcomes?

Current Study: Context

- Public sector organization.

- AC for promotion decisions.

- History of past discrimination.

- Candidates and public very concerned about fairness.

Current Study: Who

- Candidates
  - 197 total
  - Racial diversity
    - 73 White/Caucasian
    - 70 Black/African American
    - 42 Other minority
  - Gender diversity
    - 153 Males
    - 32 Females

- Assessors
  - 49 total
  - Racial diversity
    - 20 White/Caucasian
    - 17 Black/African American
    - 12 Other minority
  - Gender diversity
    - 38 Males
    - 11 Females

Current Study: What

- 3 exercises:
    - Personnel
    - Tactical
    - In-basket
- 6 dimensions:
    - Written communication, oral communication, customer service, problem solving, conflict resolution, leadership/teamwork.
- Also a written job knowledge test.

Current Study: How

- Two assessors rated each exercise.

- Assessors rotated across exercises.

  - Each candidate observed by 6 assessors total.

- Major disagreements (> 1 point) on ratings of dimensions after each exercise were addressed through discussion, then…

- Arithmetical combination of ratings by assessors, across exercises, across dimensions

Analysis Challenge

- Data set is complex!
- Ratings aren't independent:
  - Each assessor rated several candidates.
  - And each candidate was rated by two assessors; sometimes the same or sometimes different assessors
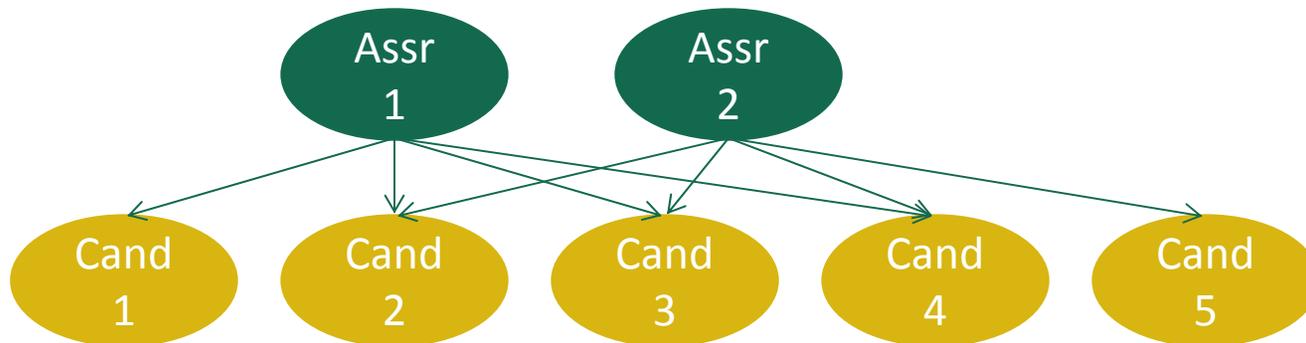
- Need to take these dependencies into account.
  - Or we may over- or under-estimate key effects.

Solution

- Solution: multilevel modeling.
  - Specifically: multivariate, cross-classified multilevel model.
- Allows us to model the complex structure of the data…
- … but interpret (largely) in regression terms.
- Analyzed each exercise separately.
  - Different assessor/candidate combinations in each exercise!

Looking for…

- How much of the variance in AC ratings is due to the assessor, rather than the candidate?
  - Does it make a difference which assessor you have?
- Main effects for candidate race and gender.
- Main effects for assessor race and gender.
- Interactions between candidate and assessor race, gender.
  - Do different assessors respond to candidate race or gender differently?
  - If so, does assessor race explain these differences?

Results: Oral Presentation Exercise - Race

- Variance explained by:
  - Candidates: 70%
  - **Assessors: 1%**
  - Residual (other factors): 29%
- Main effects:
  - Candidate race: NO
  - Assessor race: NO
- Interactions:
  - Candidate race by assessor? NO
  - Assessor race by candidate? NO

Results: Oral Presentation Exercise - Gender

- Variance explained by:
  - Candidates: 70%
  - **Assessors: 1%**
  - Residual (other factors): 29%
- Main effects:
  - Candidate gender: YES – women candidates rated higher.
  - Assessor gender: NO
- Interactions:
  - Candidate gender by assessor? NO
  - Assessor gender by candidate? NO

Results: Tactical Exercise - Race

- Variance explained by:
  - Candidates: 67%
  - **Assessors: < 1%**
  - Residual (other factors): 33%
- Main effects:
  - Candidate race: NO
  - Assessor race: NO
- Interactions:
  - Candidate race by assessor? NO
  - Assessor race by candidate? NO

Results: Tactical Exercise - Gender

- Variance explained by:
  - Candidates: 67%
  - **Assessors: 1%**
  - Residual (other factors): 32%
- Main effects:
  - Candidate gender: YES – women candidates rated higher.
  - Assessor gender: NO
- Interactions:
  - Candidate gender by assessor? NO
  - Assessor gender by candidate? NO

Results: In-Basket - Race

- Variance explained by:
  - Candidates: 75%
  - **Assessors: < 1%**
  - Residual (other factors): 24%
- Main effects:
  - Candidate race: NO
  - Assessor race: NO
- Interactions:
  - Candidate race by assessor? YES
  - Assessor race by candidate? NO
  - Candidate race had different effects for different assessors!
  - But these effects were not predicted by assessor race.
  - Overall effect is very small.

Results: In-Basket - Gender

- Variance explained by:
  - Candidates: 75%
  - **Assessors: < 1%**
  - Residual (other factors): 25%
- Main effects:
  - Candidate gender: YES – women candidates rated higher.
  - Assessor gender: NO
- Interactions:
  - Candidate gender by assessor? NO
  - Assessor gender by candidate? YES
  - Assessor gender had different effects for different candidates.
  - Women assessors gave lower ratings to women candidates – again this effect is very small.

Summary

- Most of the variance in ratings came from candidates, not assessors.
    - Assessors almost interchangeable.
- Main effects for race and gender were small or absent.
- No clear pattern of assessors favoring those with the same race or gender.

Why?

- Lack of "power" in statistical analyses?
  - Possibly.
- Good-sized sample for an AC:
  - Nearly 200 candidates and 50 assessors.
  - In multilevel modeling, number of groups is the important thing.
- However, this is a complex model.
  - No straightforward procedure for power analyses for multivariate cross-classified models.

On the other hand…

- Results were very consistent across exercises.
- The effects that were estimated were tiny.
- Results are consistent with other research on lack of similarity bias in performance appraisal.

- Reasonable to conclude that the overall trends make sense.
  - Even if we view results with a little caution.

Why?

- AC implementation?
  - Substantial training
  - Clear standards
  - Immediate accountability for ratings
- This AC follows the *Guidelines* and other best practices closely.
  - Seems to work!
  - Assessors give very similar ratings, potential sources of bias investigated here don't have an effect.

Caveats

- We can't conclude that similarity bias doesn't happen at all in ACs.
    - Looked at one sample, specific kinds of bias.
- Not all ACs follow the *Guidelines* this closely.
    - If assessor training, etc. are responsible for minimizing assessor variance, bias may play a greater role in ACs where less training, etc. is provided.

Conclusions

- The potential for ACs to be affected by demographic similarity biases is an important issue.
  - ACs are inherently judgmental and may appear to be subjective.
  - Need to address candidate and public concerns.
- This study offers evidence that ACs *can* minimize this type of bias…
  - … although we can't guarantee that for all ACs.
- Key features of the AC method seem to work as intended to minimize individual assessor effects.

Any questions?

Thank you.

For additional details, contact
Alyssa.Gibbons@colostate.edu