

# Inconsistency in Assessment Center Performance: Measurement Error or Something More?

**Alyssa Mitchell Gibbons**

**Deborah E. Rupp**

University of Illinois at Urbana-Champaign

**Deidra J. Schleicher**

Krannert School of Management - Purdue  
University

**International Congress on Assessment Center  
Methods**

11-13 October 2006

# Overview

---

- The problem of inconsistency - a brief review
- What if inconsistency isn't just error?
- Two studies:
  - Measuring inconsistency
  - Inconsistency in operational assessment centers
- Implications & future directions

# Inconsistency in Assessment Centers

- Most assessment centers measure the same dimensions across multiple exercises.
- Assume exercises behave like items on a test:

$$\begin{array}{|c|} \hline \text{observed} \\ \text{performance:} \\ \text{dimension x} \\ \text{exercise y} \\ \hline \end{array} = \begin{array}{|c|} \hline \text{true} \\ \text{performance} \\ \text{dimension x} \\ \hline \end{array} + \begin{array}{|c|} \hline \text{measurement} \\ \text{error} \\ \text{exercise y} \\ \hline \end{array}$$

- Expect that candidates should receive the same rating on the same dimension across exercises.

# However...

---

- Research repeatedly finds that candidates are *inconsistent* across exercises, e.g.:
  - Archambeau, 1979
  - Bycio, Alvares, & Hahn, 1987
  - Chan, 1996
  - Harris, Becker, & Smith, 1993
  - Highhouse & Harris, 1993
  - Jansen & Stoop, 2001
  - Joyce, Thayer, & Pond, 1994
  - Kudisch, Ladd, & Dobbins, 1997
  - Lance, Newbolt, Gatewood, Foster, French, & Smith, 2000
  - Lance, Lambert, Gewin, Lievens, & Conway, 2004
  - Neidig, Martin, & Yates, 1979
  - Sackett & Dreher, 1982
  - Schneider & Schmitt, 1992
  - Turnage & Muchinsky, 1982
  - Etc...

# Explanations for Inconsistency

---

- Assessor error
  - Cognitive load is too much – can't distinguish dimensions (*e.g., Gaugler & Thornton, 1989; Lance, Foster et al., 2004*)
  - Common rater effects (*e.g., Kolk, Born, & van der Flier, 2002*)
- Rating process
  - Focus: within-exercise or within-dimension (*Robie et al., 2000*)
- Exercise design
  - Trait activation theory (*Haaland & Christiansen, 2002*)
- Attempts to alleviate error have had mixed **SUCCESS**. (*Lievens, 2001a; Lance et al., 2004; cf Arthur,*

# Real Inconsistency?

---

- Lievens (2001; 2002) videotape studies:
  - Assessors' ratings pick up on consistent, differentiated performance *when it is there*.
- Growing evidence that candidates really do perform differently in different exercises. (*Lance et al., 2004*)
  - Are exercises just independent work samples?
- Consider: all this research is focused on the population of candidates as a whole.
  - Either consistent across exercises, differentiated across dimensions
  - Or inconsistent across exercises, undifferentiated across dimensions

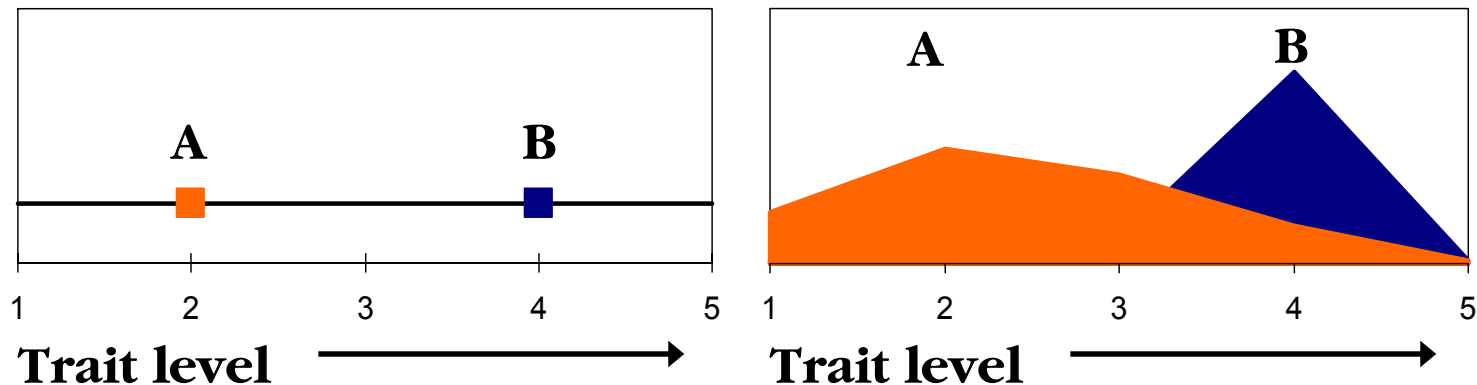
# Inconsistency in Other Domains

---

- Person-situation debate in personality psychology (*Mischel, 1968, etc.*).
  - Is behavior determined by person factors (dimensions)?
  - Or situational influences (exercises)?
- Increasingly, the answer is **both**. (*e.g., Mischel, 2004*)
- One partial resolution:
  - Some people may be more consistent than others. (*Bem & Allen, 1974; Krueger et al., 1996; Lord, 1982; Turner & Gilliam, 1979; Zuckerman, Bernieri, Koestner, & Rosenthal, 1989*)

# Consistency as an Individual Difference

- Fleeson (2001) argued that personality-trait-related behavior is best expressed as a **distribution**:

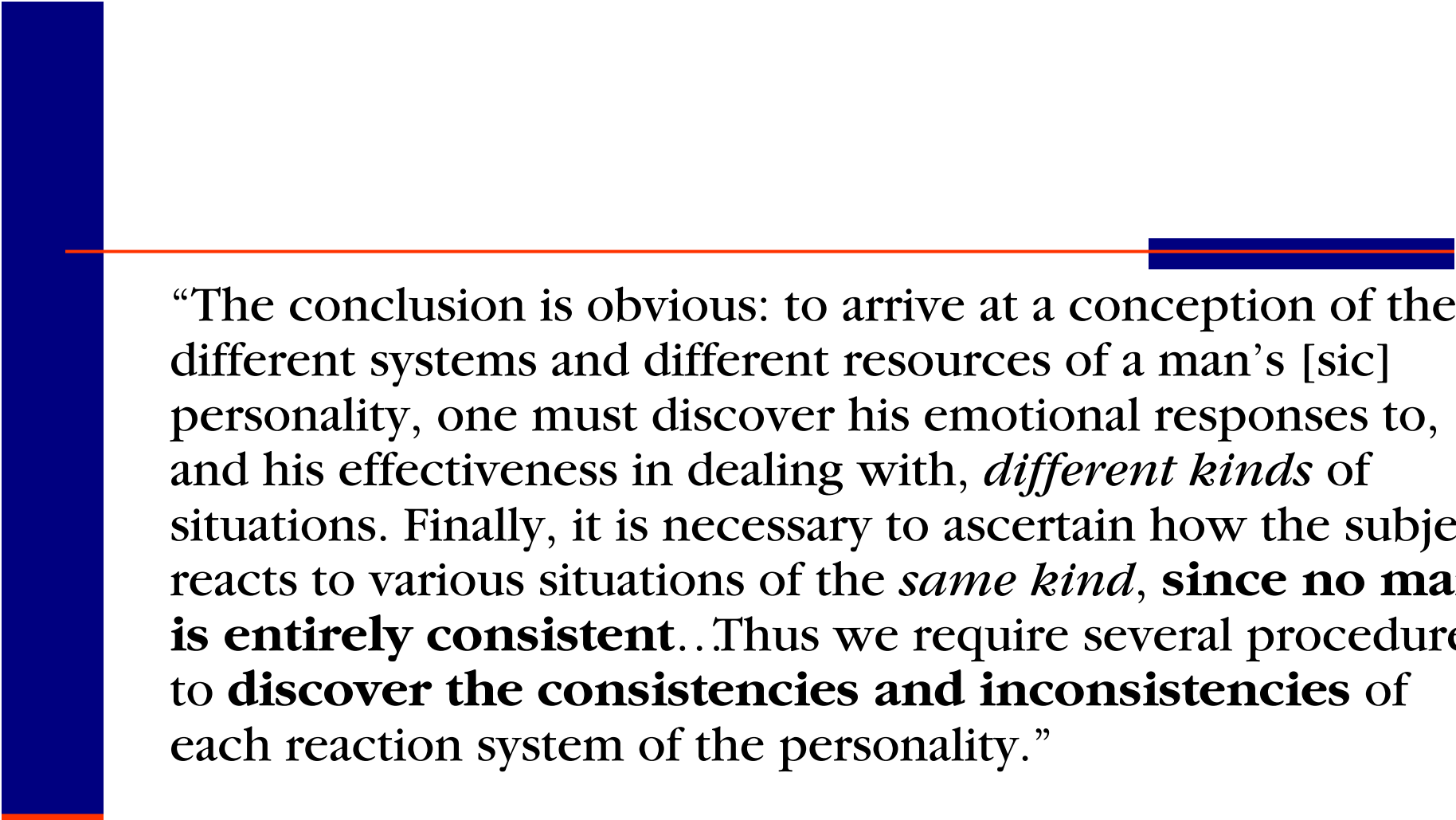


- Found that the variance of behavior was a reliable individual difference.

# An Interesting Question

---

- If the inconsistency measured in assessment centers is not just assessor error...
- And if inconsistency may be an individual difference between persons...
- **Might the inconsistency we see in assessment centers tell us something important about the candidates?**



“The conclusion is obvious: to arrive at a conception of the different systems and different resources of a man’s [sic] personality, one must discover his emotional responses to, and his effectiveness in dealing with, *different kinds* of situations. Finally, it is necessary to ascertain how the subject reacts to various situations of the *same kind*, **since no man is entirely consistent**. . . Thus we require several procedures to **discover the consistencies and inconsistencies** of each reaction system of the personality.”

*Office of Strategic Services Assessment Team  
Fiske, Hanfmann, MacKinnon, Miller, & Murray, 1948*

# Does Consistency Matter?

- Consistent high performers are obviously desirable employees...what about the people in the middle?

- Consistent mediocre performer vs. highly variable performer?

- If you were a college basketball coach...

Game:	1	2	3	4	Avg
Consistent player	10	11	10	9	<b>10</b>
Inconsistent player	2	16	18	4	<b>10</b>

# Does Consistency Matter? (cont.)

---

- In college basketball, consistency of scoring incrementally predicts team performance over & above individual average scoring (*Gibbons & Rupp, 2006*).
  - Our consistent player's predicted win %: 64%
  - Our inconsistent player's predicted win %: 45%
- Management Progress Study: consistency-like factors related to job outcomes (*Bray, Campbell, & Grant 1974*).
  - Behavioral flexibility
  - Stability of performance

# Implications

---

- If consistency in AC performance is an individual difference...
- ..and if it predicts organizationally relevant outcomes...
- ..we may be losing valuable information when we treat inconsistency as measurement error.

# Two Studies

---

- Can we measure consistency in an AC context?
  - Propose individual-level indices of consistency & differentiation across dimensions.
- Are there individual differences in consistency in operational ACs? Are those differences meaningful?
  - Apply the indices to data from three operational assessment centers.

# Measuring Consistency

---

- Need an individual-level measure of consistency
- Previous studies of consistency look at within-person variance, but...
  - Many observations (30+) are needed to estimate variance reliably. (*Fleeson, 2001; Gibbons & Rupp, 2006*)
  - Assessment centers have multiple observations, but not that many.
  - Unlikely to find stable estimates of within-dimension variance at the individual level.

# Proposed Indices

---

- Propose two indices (consistency and differentiation) that can be calculated at the individual level.
- Based on individual matrix of within-exercise dimension ratings.
- Main idea: compare consistency across pairs of exercises to gain a picture of overall consistency.
  - How often does a candidate receive the same rating on the same dimension in 2 different exercises?
- Differentiation measures the extent to which the candidate's performance is the same across dimensions.
  - How often does a candidate receive the same rating on 2 different dimensions measured in the same exercise?

# Numerical Example: Consistency

Consider the matrix of ratings for hypothetical candidate Ron:

RON	Exercises				
Dimensions:	1	2	3	4	5
1	4	5	5	4	5
2	3	2	3	3	2
3	2	2	1	2	1
4	1	1	2	2	1
5	5	4	4	4	5

Compare one pair of exercises at a time. Did our candidate receive essentially the same pattern of ratings in both exercises?

Rating scale: 1 -5

**Pairwise consistency index:** root mean squared discrepancy between ratings of the same dimension in two exercises.

$$c_{12} = \sqrt{\frac{(4-5)^2 + (3-2)^2 + (2-2)^2 + (1-1)^2 + (5-4)^2}{5}} = .77$$

# Numerical Example: Consistency

- Repeat the process for all possible pairs of exercises:

$$c_{13} = .89 \quad c_{14} = .63 \quad c_{15} = .77 \quad c_{23} = .77$$

$$c_{24} = .77 \quad c_{25} = .77 \quad c_{34} = .63 \quad c_{35} = .77 \quad c_{45} = 1.00$$

- Find the average of the pairwise indices  $c_{ik}$ :

$$= .77$$

- For interpretation, divide by the range of the rating scale:

$$.77/4 = .19$$

- Subtract from 1 so that higher values = greater consistency:

# Numerical Example: Differentiation

Consider the matrix of ratings for hypothetical candidate Ron:

RON	Exercises				
Dimensions:	1	2	3	4	5
1	4	5	5	4	5
2	3	2	3	3	2
3	2	2	1	2	1
4	1	1	2	2	1
5	5	4	4	4	5

Compare one pair of dimensions at a time. Did our candidate receive different ratings on different dimensions?

Rating scale: 1 -5

**Pairwise differentiation index:** root mean squared discrepancy between ratings of a pair of dimensions in the same exercise (across all exercises).

$$d_{12} = \sqrt{\frac{(4-3)^2 + (5-2)^2 + (5-3)^2 + (4-3)^2 + (5-2)^2}{5}} = 2.19$$

# Numerical Example: Differentiation

- Repeat the process for all possible pairs of dimensions:

$$d_{13} = .89 \quad d_{14} = 3.28 \quad d_{15} = .77 \quad d_{23} = 1.18 \quad d_{24} = 1.26$$

$$d_{25} = 1.95 \quad d_{34} = .77 \quad d_{35} = 2.90 \quad d_{45} = 3.13$$

- Find the average of the pairwise indices  $d_{ij}$ :

$$= 2.06$$

- For interpretation, divide by the range of the rating scale:

$$2.06/4 = D = .51$$

- Higher values = greater differentiation. Do not subtract from 1.

# Interpretation

- Ron's index values:
  - Consistency: .81
  - Differentiation: .51
- Ron was highly consistent across exercises and moderately differentiated across dimensions.

This seems appropriate, given Ron's ratings.

RON	Exercises				
Dimensions:	1	2	3	4	5
1	4	5	5	4	5
2	3	2	3	3	2
3	2	2	1	2	1
4	1	1	2	2	1
5	5	4	4	4	5

# Comparing Possible Profiles

Ron		High C / High				
D	E1	E2	E3	E4	E5	
D1	4	5	5	4	5	
D2	3	2	3	3	2	
D3	2	2	1	2	1	
D4	1	1	2	2	1	
D5	5	4	4	4	5	

**C = .81**  
**D = .51**

Fred		Low C / High				
D	E1	E2	E3	E4	E5	
D1	1	1	3	4	1	
D2	3	2	5	5	4	
D3	2	1	1	5	2	
D4	3	1	4	5	3	
D5	2	3	3	5	3	

**C = .52**  
**D = .34**

Hermione		High C / Low				
D	E1	E2	E3	E4	E5	
D1	2	3	3	3	3	
D2	3	3	3	2	3	
D3	3	3	4	3	3	
D4	3	4	3	3	3	
D5	3	3	3	3	3	

**C = .86**  
**D = .14**

George		Low C / Low D				
	E1	E2	E3	E4	E5	
D1	5	4	3	2	1	
D2	5	4	3	2	1	
D3	5	4	3	2	1	
D4	5	4	3	2	1	
D5	5	4	3	2	1	

**C = .50**  
**D = .00**

# Computer Simulation of C and D

---

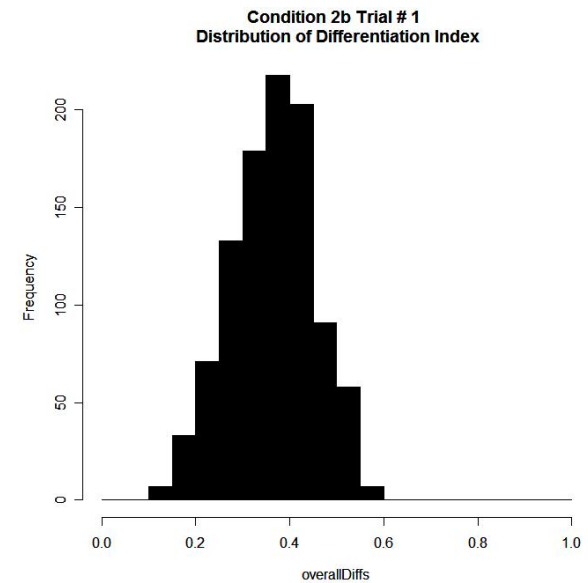
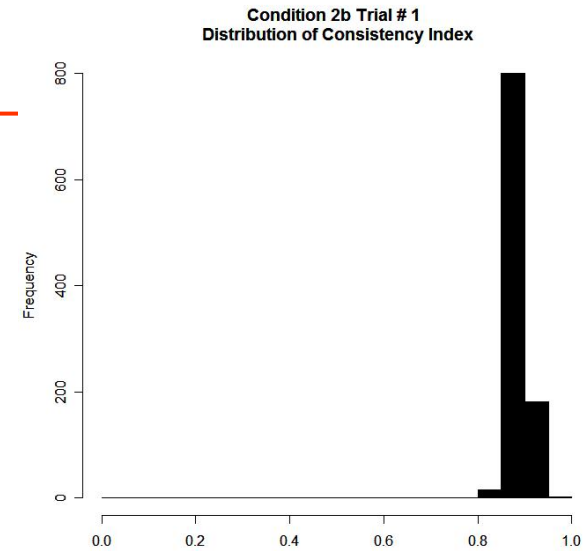
- Are the *C* and *D* indices useful for identifying individual differences in AC candidates?
- Must show reliability - indicate differences when and only when they exist.
- Simulate samples of hypothetical AC candidates using various assumptions, e.g.:
  - Everyone is consistent & differentiated
  - Everyone is inconsistent & undifferentiated
  - Candidates vary in consistency and/or differentiation

# Details

- Used R programming system (*R Development Core Team, 2005*)
- Generated individual matrix of exercise x dimension ratings for each candidate.
  - Consistent: constrained to be within +/- 1 rating point across exercises
  - Differentiated: constrained to be within +/- 1 rating point across dimensions
  - Inconsistent/undifferentiated: random
- Focus is on distributions of  $C$  &  $D$  indices, split-half reliability based on  $c_{ij}$  and  $d_{ij}$

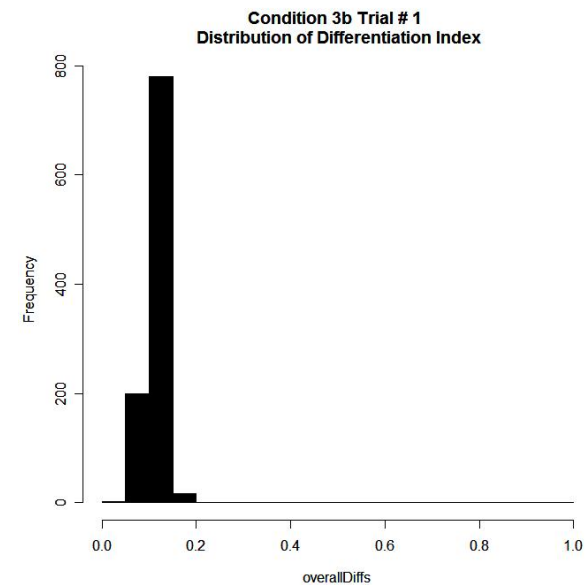
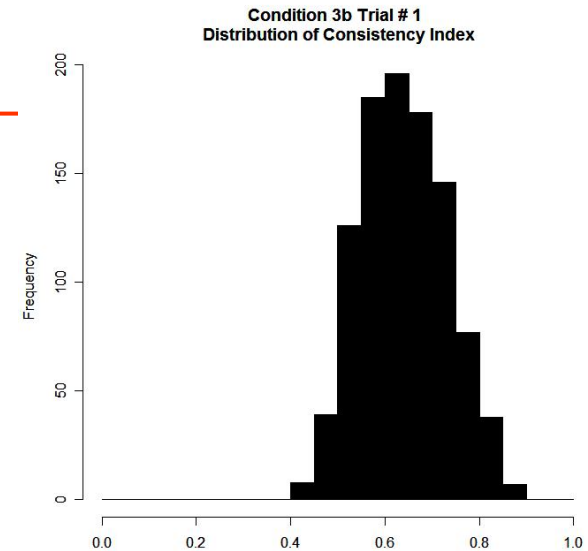
# Example 1

- All candidates consistent & differentiated.
  - 5 dimensions
  - 5 exercises
  - Rating scale: 1-7
  - $N = 1000$
- Avg  $C = .89$
- Avg  $D = .36$
- Split-half  $r(C) = .39$
- Split-half  $r(D) = .27$



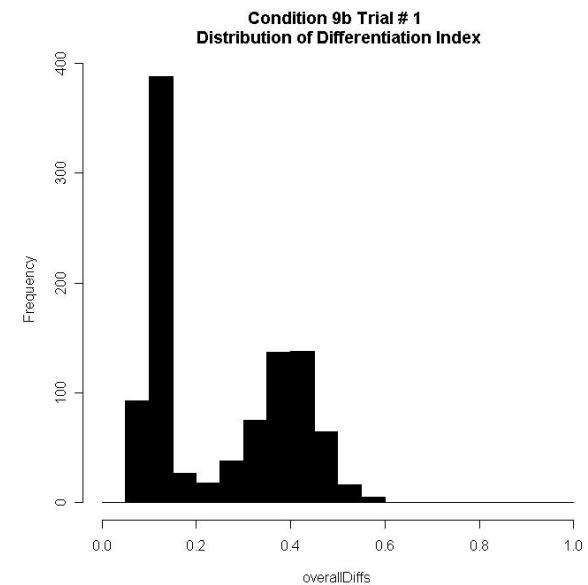
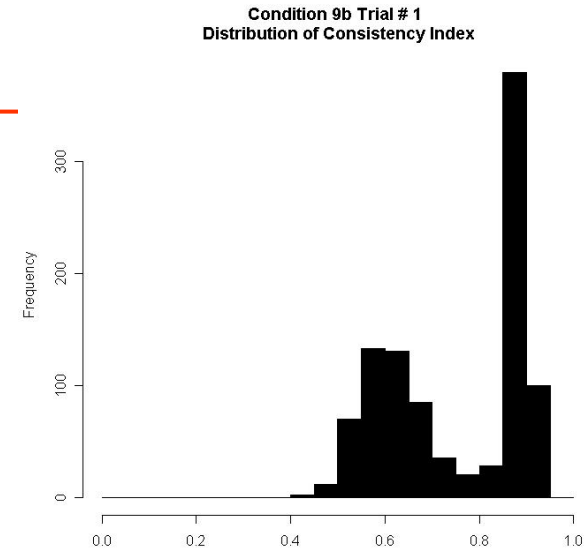
# Example 2

- All candidates *inconsistent* & *undifferentiated*.
  - 5 dimensions
  - 5 exercises
  - Rating scale: 1-7
  - N = 1000
- Avg  $C = .64$
- Avg  $D = .11$
- Split-half  $r(C) = .28$
- Split-half  $r(D) = .39$



# Example 3

- Candidates vary in consistency and differentiation.
  - 5 dimensions
  - 5 exercises
  - Rating scale: 1-7
  - $N = 1000$
- Avg  $C = .75$
- Avg  $D = .25$
- Split-half  $r(C) = .87$
- Split-half  $r(D) = .87$



# Computer Simulation Summary

---

- Over a variety of conditions:
- When there **is** variation in consistency and/or differentiation:
  - Avg. split-half  $r(C) = .81$
  - Avg. split-half  $r(D) = .81$
- When there is **not** variation in consistency and/or differentiation:
  - Avg. split-half  $r(C) = .51$
  - Avg. split-half  $r(D) = .51$

# C and D in Operational ACs #1 & 2

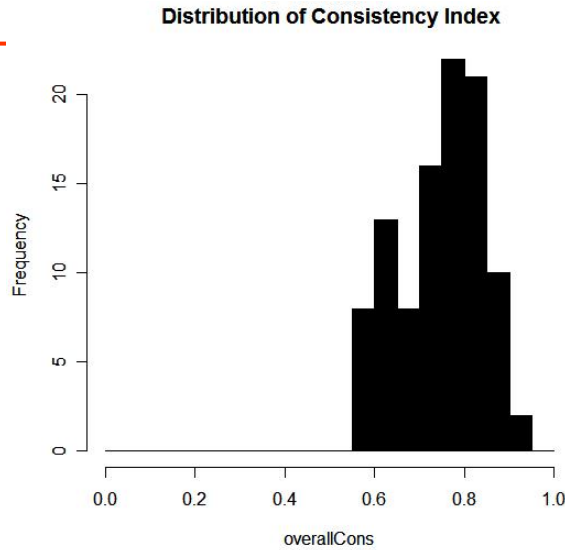
---

- How are the C and D indices distributed in real groups of assessment center candidates?
- 2 ACs - civil service promotional program
  - N = 100 & N = 197
  - 3 exercises, 6 dimensions, rating scale 1-5
  - 2 assessors per exercise
  - Focus on distribution, interrater reliability

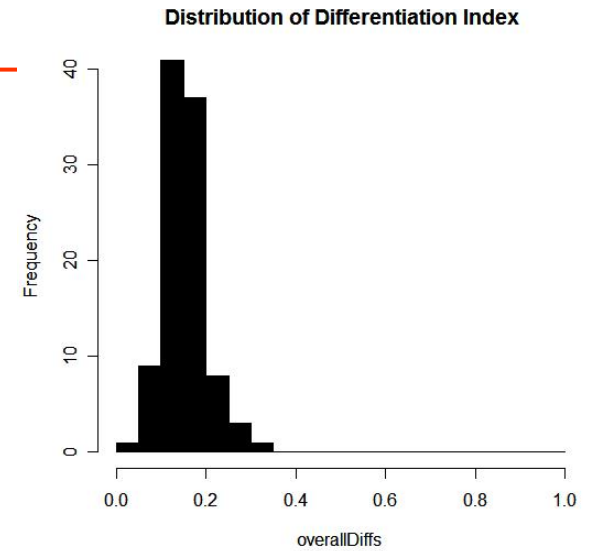
# C and D Distributions

**AC #1:**

**Avg. C  
= .74**

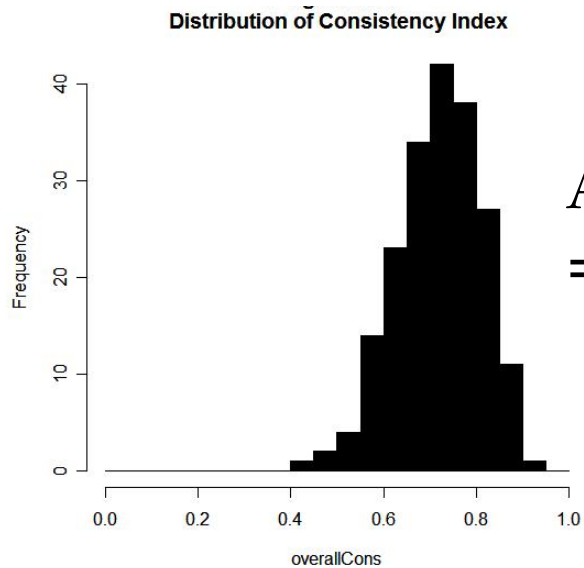


**Avg. D  
= .15**

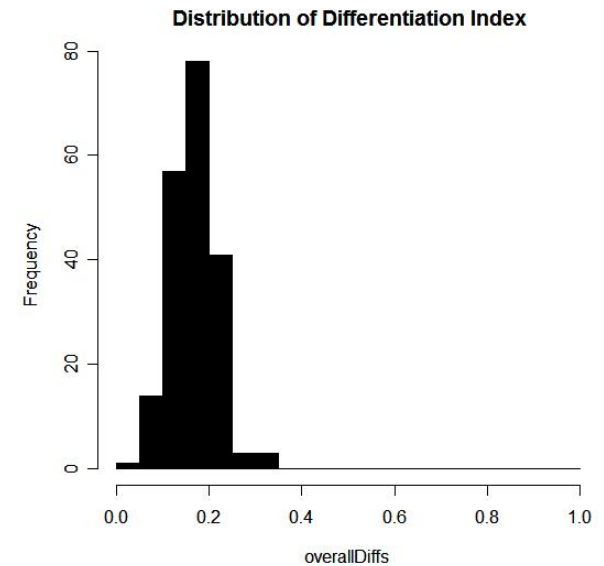


**AC #2:**

**Avg. C  
= .72**



**Avg. D  
= .17**



# C and D in Operational AC #3

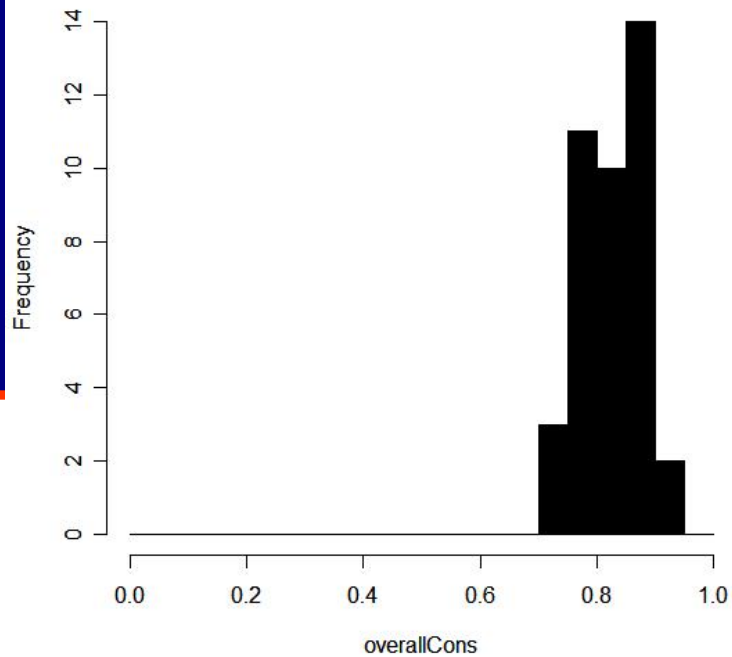
---

- Two unanswered questions:
  - What about assessor effects?
  - Do the C and D indices tell us anything meaningful about the candidates?
- Reanalysis of student assessment center data  
*(Schleicher, Day, Mayes, & Riggio, 2002)*
  - Expert true scores + multiple assessor ratings.
  - Outcome data - supervisors' performance ratings, independent of the AC.
  - N = 40 (N = 27 for supervisor ratings)
  - 3 exercises, 3 dimensions, rating scale 1-7

# C and D Distributions - AC #3

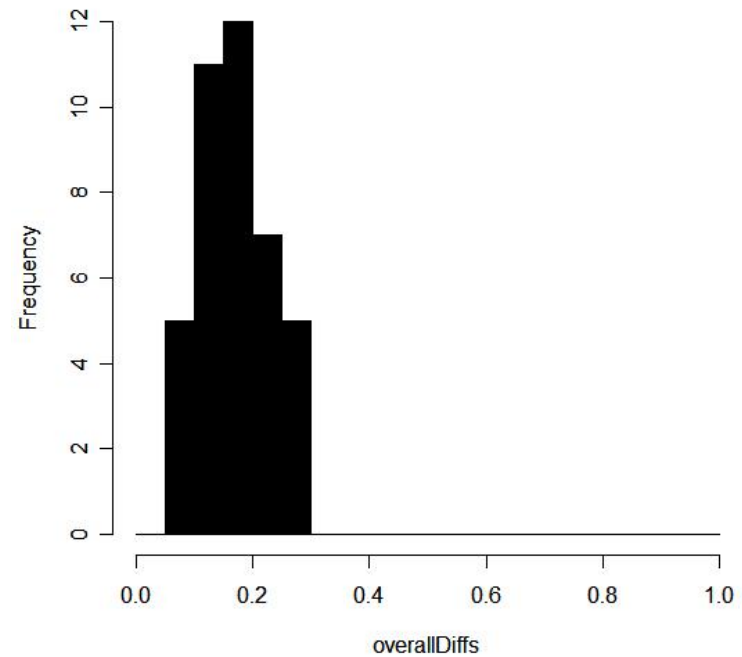
Avg. C =  
.83

Expert True Scores  
Distribution of Consistency Index



Avg. D = .17

Expert True Scores  
Distribution of Differentiation Index



# C and D Distributions - AC # 3

## (cont.)

---

- Data doesn't exactly fit any of the simulated models - appears to be somewhere in between:
  - a model where candidates vary in both C and D.
  - a model where all candidates are consistent and undifferentiated.
  - Some variation in C and D, but not a lot.
- C and D indices based on assessor ratings correlate poorly with true-score C and with each other.
  - Assessors were students with 90 minutes of training.

# C, D, and Job Outcomes

- True-score C & D indices correlated with supervisor performance ratings:
  - $r(C, \text{perf})$ : .42
  - $r(D, \text{perf})$ : -.39
- C index predicts supervisor ratings **over and above** overall assessment rating:
  - OAR alone:  $R = .46$ ,  $R^2 = .21$ ,  $p < .05$
  - OAR + C:  $R = .58$ ,  $R^2 = .34$ ,  $p < .05$
  - $\Delta R^2 = .13$ ,  $p < .05$
- In other words... knowing consistency substantially improves our predictive power!
- Given two candidates with the same OAR... the more consistent candidate is a better employee two years later.

# Conclusions

---

- Findings are preliminary, but...
- *C* and *D* appear to be useful, reliable.
- Real assessment center participants vary in consistency.
- Consistency is related to job outcomes.
  - Even when average AC performance is controlled.
  - Higher consistency leads to higher on-the-job performance ratings.
- Inconsistency is more than just measurement error or an inherent quirk in the assessment center process.

# Caveats

---

- Doesn't *eliminate* problem of assessor error.
  - AC #3 highlights the need for good training.
  - Other design features matter.
- Needs replication!
  - Larger samples, other types of ACs.
- Not ready for use in selection (yet).
  - Not reliable enough for personnel decisions.
  - Must demonstrate relevance for target job.

# Implications

---

- If consistency in AC performance is an individual difference...
- ..and if it predicts organizationally relevant outcomes...
- ..we **are** losing valuable information when we treat inconsistency as measurement error.
- Understanding consistency can help us make more effective use of assessment center data.
  - Selection/promotion (someday- ?)
  - Performance management
  - And especially...

# Consistency and Development

Inconsistent  
average  
performer



Can excel  
but doesn't  
always

Consistent  
average  
performer



Doesn't fail  
but can't  
excel

- How do we develop these two employees?
- Applying effective behaviors more consistently vs. learning new effective behaviors.
- Provide feedback based on **patterns** of behavior.

# Future Directions

---

- When does consistency matter?
  - Ad exec vs. air traffic controller.
- What drives consistency?
  - Motivation? (*Kane, 1986*)
  - Personality? Self-monitoring? (*Kuypsch, Kleinmann, & Koller, 1998*)
  - Adaptability? (*Pulakos, Arad, Donovan, & Plamondon, 2000*)
  - Differences in skills? (*Mischel, 1984*)



Questions?



Thank You!