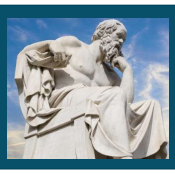
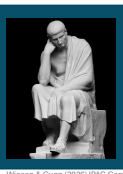
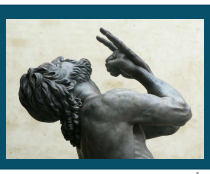
Two-Board Grading Using BARS:

Attempts to Understand Disturbing Reliability Findings

Joel P. Wiesen - *Ph.D., Director, Applied Personnel Research*
Joel J. Gunn - *MBA, Personnel Analyst III, City of Columbus Civil Service Commission*

1

Overview

- Fundamental Assumption
  High rater agreement indicates valid, accurate grading
- **Will present new data that disputes that assumption**
- **Two grading boards each grade all candidates**
- **High reliability within boards**
- **Surprisingly low correlation between boards**

2

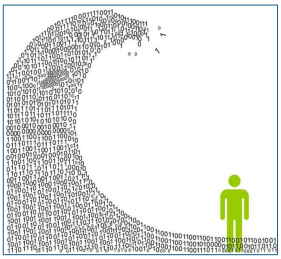
Short History

- Wiesen (2025 IPAC) – The BARS Inter-rater Reliability Illusion
  - Initial Finding: low between board correlation
  - 12 situational exercises graded using BARS
  - Expected between board correlation was higher than observed
- Wiesen (2026 SIOP) – Reliability of BARS: New Data from Two-Board Grading
  - Initial attempt to understand
  - Detailed Venn Diagram
- Wiesen & Gunn (2026 IPAC) – Today's update
  - 66 situational exercises, 5 job titles
  - More data analysis to attempt to understand
  - Guide for practice

3

Question #1
Is lower inter-board correlation a general phenomenon?

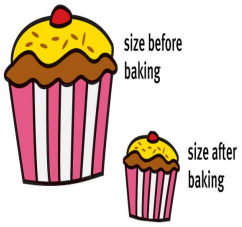
- We gathered data from;
  - 5 job titles
    - Police Sergeant
    - Police Lieutenant
    - Police Commander
    - Fire Lieutenant
    - Fire Captain
  - 12 years of exams
  - 66 total exercises scored using BARS



4

Question #1
Is lower inter-board correlation a general phenomenon?

- Define shrinkage
- Five step statistical analysis
  - Spearman-Brown for Board 1 & Board 2
  - Observed correlation between boards
  - Expected correlation between boards
  - Percent *(observed / expected)*
  - Shrinkage *(1 – percent)*



5

Question #1
Is lower inter-board correlation a general phenomenon?

Police Sergeant

Year	Exercise	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/expt)	Shrinkage
2013	1	0.92	0.92	0.92	0.80	87%	13%
2013	2	0.89	0.86	0.87	0.65	74%	26%
2015	1	0.89	0.95	0.92	0.48	52%	48%
2015	2	0.89	0.87	0.88	0.77	88%	12%
2017	1	0.88	0.80	0.84	0.76	91%	9%
2017	2	0.84	0.88	0.86	0.66	77%	23%
Mean (2013-2025)		0.91	0.90	0.91	0.63	70%	30%

Minimum Observed Correlation **0.44**

6

Question #1
Is lower inter-board correlation a general phenomenon?

Police Lieutenant

Year	Exercise	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/expt)	Shrinkage
2012	1	0.91	0.94	0.92	0.81	87%	13%
2012	2	0.81	0.94	0.87	0.70	81%	19%
2014	1	0.91	0.90	0.90	0.68	75%	25%
2014	2	0.96	0.96	0.96	0.86	89%	11%
2016	1	0.90	0.93	0.91	0.75	82%	18%
2016	2	0.92	0.84	0.88	0.42	48%	52%
Mean (2012-2024)		0.91	0.92	0.91	0.69	75%	25%

Minimum Observed Correlation **0.36**

7

Question #1
Is lower inter-board correlation a general phenomenon?

Police Commander

Year	Exercise	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/exp	Shrinkage
2012	1	0.95	0.90	0.92	0.93	101%	-1%
2012	2	0.94	0.93	0.93	0.62	66%	34%
2014	1	0.88	0.89	0.88	0.80	91%	9%
2014	2	0.81	0.88	0.85	0.92	108%	-8%
2016	1	0.80	0.82	0.81	0.42	52%	48%
2016	2	0.91	0.89	0.90	0.76	84%	16%
Mean (2012-2024)		0.85	0.87	0.85	0.69	80%	20%

Minimum Observed Correlation **0.42**

8

Question #1
Is lower inter-board correlation a general phenomenon?

Fire Lieutenant

Year	Exercise	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/expt)	Shrinkage
2013	1	0.94	0.93	0.93	0.77	82%	18%
2013	2	0.92	0.90	0.91	0.65	71%	29%
2015	1	0.95	0.94	0.94	0.81	86%	14%
2015	2	0.95	0.86	0.90	0.58	64%	36%
2017	1	0.94	0.93	0.93	0.81	87%	13%
2017	2	0.94	0.96	0.95	0.61	64%	36%
Mean (2013-2023)		0.91	0.91	0.91	0.71	78%	22%

Minimum Observed Correlation **0.58**

9

Question #1
Is lower inter-board correlation a general phenomenon?

Fire Captain

Year	Exercise	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/expt)	Shrinkage
2013	1	0.86	0.92	0.89	0.57	64%	36%
2013	2	0.90	0.95	0.92	0.61	66%	34%
2015	1	0.92	0.89	0.90	0.65	72%	28%
2015	2	0.91	0.95	0.93	0.76	82%	18%
2017	1	0.83	0.58	0.69	0.47	68%	32%
2017	2	0.85	0.90	0.87	0.64	73%	27%
Mean (2013-2023)		0.87	0.86	0.86	0.66	76%	24%

Minimum Observed Correlation **0.47**

10

Question #1
Is lower inter-board correlation a general phenomenon?

Summary – All Ranks

Rank	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/expt)	Shrinkage
Police Sergeant	0.91	0.90	0.91	0.63	70%	30%
Police Lieutenant	0.91	0.92	0.91	0.69	75%	25%
Police Commander	0.85	0.87	0.85	0.69	80%	20%
Fire Lieutenant	0.91	0.91	0.91	0.71	78%	22%
Fire Captain	0.87	0.86	0.86	0.66	76%	24%
Mean of Means	0.89	0.89	0.89	0.67	76%	24%

11

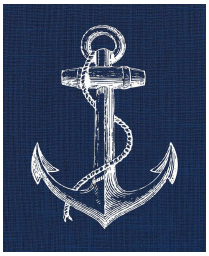
Question #1
Is lower inter-board correlation a general phenomenon?

- Yes, lower inter-board correlation is a general phenomenon.

12

Question #2
Does the number of anchors on the BARS relate to shrinkage?

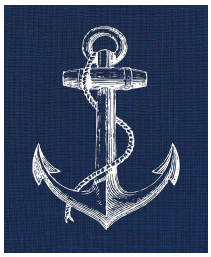
- Background of testing program
  - 66 situational exercises,132 rater boards
  - 4 dimensions per exercise
  - Total number of anchors per dimension
  - Total number of anchors per exam
  - Focus on exercises where Observed Correlation is less than 0.50
    - *High shrinkage*



13

Question #2
Does the number of anchors on the BARS relate to shrinkage?

- Police Sergeant data example *(next slide)*
- Mean number of anchors = 89
- No pattern below, at or above the mean number of anchors
  - Likewise for other job titles



14

Question #2
Does the number of anchors on the BARS relate to shrinkage?

Police Sergeant

Year	Exercise	BARS Total	N	Mean	Std.Dev	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/expt)	Shrinkage
2015	1	98	109	47.84	8.42	0.89	0.95	0.92	0.48	52%	48%
2019	1	67	124	55.20	10.70	0.97	0.91	0.94	0.44	47%	53%
2021	1	69	104	57.77	5.01	0.86	0.87	0.86	0.48	55%	45%
2025	1	108	80	43.70	7.05	0.92	0.93	0.92	0.48	52%	48%
	MEAN	89.4	92.71	50.65	8.20	0.91	0.90	0.91	0.63	70%	30%

15

Question #2
Does the number of anchors on the BARS relate to shrinkage?

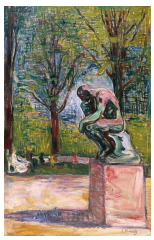
- No relationship with # of anchors

NEXT »

16

Question #3
Does subject matter predict shrinkage?

- Each of the 66 exercises had a topic.
  - Technical / tactical
  - Interpersonal
  - Procedural
  - Inflammatory
  - Are some subjects more difficult to score?
  - Does this affect shrinkage?



17

Question #3
Does subject matter predict shrinkage?

- Reviewed each exercise
- Noted subject matter
- Grouped into categories
- Created a matrix for each rank
- Converted to binary code
- Calculated correlation with shrinkage

18

Question #3
Does subject matter predict shrinkage?

Police Sergeant Data - Example

Year	Exercise	arrest	shots fired	vehicle accident	de-escalation	admin investigation	DEI issue	protest / demonstration
2013	1	0	0	1	0	0	0	0
2013	2	0	0	0	1	0	0	0
2015	1	0	1	0	1	0	1	1
2015	2	0	0	1	0	1	0	0
2017	1	0	0	0	1	0	0	1
2017	2	0	0	0	0	1	0	0
shrinkage (2013-25)		**0.432**	**0.18**	**-0.51**	**0.30**	**-0.14**	**0.61**	**-0.05**
stdv (2013-25)		**-0.34**	**0.55**	**0.03**	**-0.55**	**-0.59**	**0.24**	**-0.16**

19

Question #3
Does subject matter predict shrinkage?

- Interpreting the data
  - Positive correlation means the use of that *subject matter category* is associated with higher shrinkage
  - Higher shrinkage is bad

20

Question #3
Does subject matter predict shrinkage?

- Interpreting the data
- Some interesting data points;
- Topic: Vehicle accident -0.51
  - Negative correlation with shrinkage
  - Subject highly procedural
  - Well documented in internal source material *(Directives, SOPs, etc.)*
  - Step-by-step instructions for supervisors

21

Question #3
Does subject matter predict shrinkage?

- Interpreting the data
- Topic: DEI Issue +0.61
  - Positive correlation with shrinkage
  - Subject is complex
  - Subject is less procedural *(subjective)*
  - Subject is often emotionally charged

22

Question #3
Does subject matter predict shrinkage?

- Counter example
- Topic: Arrest +0.43
- Topic: Crime Scene +0.46
  - Positive correlation with shrinkage
  - Subject is procedural
  - Subject is clear in sources

23

Question #3
Does subject matter predict shrinkage?

- Some indication that more specific source material results in less shrinkage
- Perhaps clarity of grading criteria is the culprit
- Worthy of further study

24

Question #4
Does subject matter predict standard deviation within a board?

Police Sergeant Data - Example

Year	Exercise	arrest	shots fired	vehicle accident	de-escalation	admin investigation	DEI issue	protest / demonstration
2013	1	0	0	1	0	0	0	0
2013	2	0	0	0	1	0	0	0
2015	1	0	1	0	1	0	1	1
2015	2	0	0	1	0	1	0	0
2017	1	0	0	0	1	0	0	1
2017	2	0	0	0	0	1	0	0
shrinkage (2013-25)		**0.432**	**0.18**	**-0.51**	**0.30**	**-0.14**	**0.61**	**-0.05**
stdv (2013-25)		**-0.34**	**0.55**	**0.03**	**-0.55**	**-0.59**	**0.24**	**-0.16**

25

Question #4
Does subject matter predict standard deviation within a board?

- Topic: Administrative -0.59 (logical)
  - Negative correlation with rater SD
  - Topic is well covered in sources
- Topic: Shots fired +0.55 (Illogical)
  - Positive correlation with rater SD
  - Topic is well covered in sources
- Topic: De-escalation -0.55 (Illogical)
  - Negative correlation with rater SD
  - Topic is subjective, less covered in sources

26

Question #4
Does subject matter predict standard deviation within a board?

- Interpreting the Sgt data
- 3 of 24 topics statistically significant
- Perhaps different PDs have different SOPs and raters base their ratings on their home PD SOPs

27

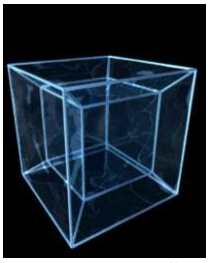
Question #4
Does subject matter predict standard deviation within a board?

- Interpreting the data
- No correlation between these two correlations:
  - SD of raters with shrinkage
  - Subject matter with shrinkage
- What other analyses make sense?

28

Question #5
Dimension Effects: Do some dimensions yield better results?

- **4 Sgt exercises with overall shrinkage greater than 0.50**
- Looked at each dimension
- Conducted the same statistical analysis
  - Spearman-Brown for Board 1 & Board 2
  - Observed & expected correlations
  - Shrinkage *(i.e., 1 – observed/expected)*



29

Question #5
Dimension Effects: Do some dimensions yield better results?

- Police Sergeant Data
  - *Dimension 1:* ***Oral Communication***

Year	Exercise	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/expt)	Shrinkage
2015	1	0.87	0.91	0.89	0.54	61%	39%
2019	1	0.95	0.81	0.88	0.36	41%	59%
2021	1	0.82	0.83	0.82	0.37	45%	55%
2025	1	0.88	0.89	0.88	0.58	65%	35%

30

Question #5
Dimension Effects: Do some dimensions yield better results?

- Police Sergeant Data
  - *Dimension 2:* ***Interpersonal Relations***

Year	Exercise	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/expt)	Shrinkage
2015	1	0.81	0.93	0.87	0.45	52%	48%
2019	1	0.95	0.76	0.85	0.43	50%	50%
2021	1	0.80	0.85	0.83	0.51	62%	38%
2025	1	0.88	0.91	0.90	0.31	35%	65%

31

Question #5
Dimension Effects: Do some dimensions yield better results?

- Police Sergeant Data
  - *Dimension 3:* ***Information Analysis***

Year	Exercise	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/expt)	Shrinkage
2015	1	0.82	0.93	0.87	0.31	36%	64%
2019	1	0.96	0.87	0.91	0.42	46%	54%
2021	1	0.72	0.81	0.76	0.39	51%	49%
2025	1	0.86	0.82	0.84	0.35	41%	59%

32

Question #5
Dimension Effects: Do some dimensions yield better results?

- Police Sergeant Data
  - *Dimension 4:* ***Problem Sensing & Resolution***

Year	Exercise	Sprmn-Brn Board 1	Sprmn-Brn Board 2	Expected Correlation	Observed Correlation	Percent (obs/expt)	Shrinkage
2015	1	0.86	0.94	0.90	0.37	41%	59%
2019	1	0.95	0.90	0.93	0.43	46%	54%
2021	1	0.82	0.85	0.84	0.27	32%	68%
2025	1	0.90	0.86	0.88	0.55	63%	37%

33

Question #5
Dimension Effects: Do some dimensions yield better results?

- No.
- FYI - Some studies indicate that SMEs have difficulty distinguishing between dimensions
  - Jackson, Michaelides, Dewberry, Nelson & Stephens (2022)
  - Jackson, Blair & Ingold (2024)
  - Jackson, Michaelides, Dewberry & Yang (2024)

34

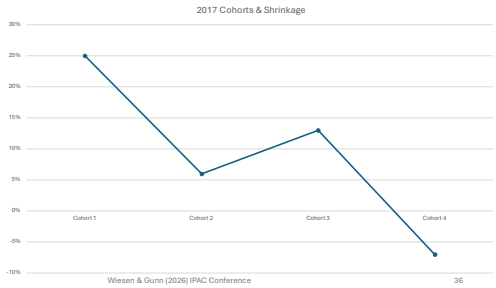
Question #6
Does the order in which candidates were scored affect shrinkage?

- Focused on 3 of the high shrinkage exercises (>.50)
- Created candidate cohorts of 15 based on the order in which they were scored
- Conducted the same statistical analysis
  - Spearman-Brown Board 1 & Board 2
  - Observed & expected correlation
  - Shrinkage
- **QUESTION**: Will shrinkage change as boards gain experience using BARS?

35

Question #6
Does the order in which candidates were scored affect shrinkage?

- Police Sergeant Data
- 2017 Exercise 1

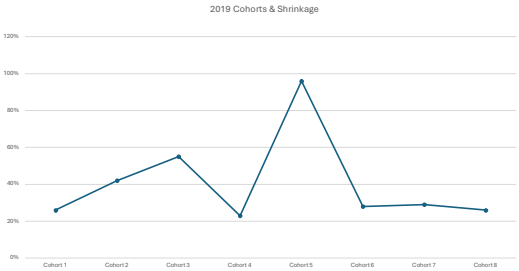


36

Question #6

Does the order in which candidates were scored affect shrinkage?

- Police Sergeant Data
- 2019 Exercise 1

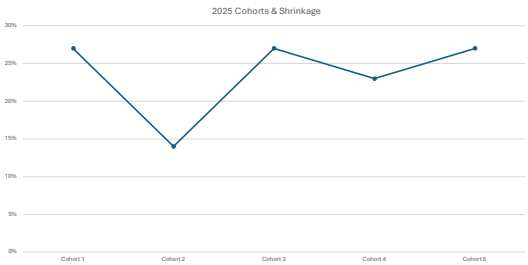


37

Question #6

Does the order in which candidates were scored affect shrinkage?

- Police Sergeant Data
- 2025 Exercise 1



38

Question #6

Does the order in which candidates were scored affect shrinkage?

- No. This was not a general finding
- This looked at overall grades, not individual dimensions

39

Question #7
Is there a correlation between sequential order and score?

- Does the order in which candidates are scored affect overall scores?
- Top 80 candidates
- Scores of candidates by rater in order of grading

40

Question #7
Is there a correlation between sequential order and score?

- Example: 2025 Police Sergeant Exam, Exercise 1, Cohort 1

Sequential #	Rater 1 Total	Rater 2 Total	Rater 3 Total	Board 1 Total	Rater 4 Total	Rater 5 Total	Rater 6 Total	Board 2 Total
1	11	13	10	34	13	7	13	33
2	32	31	28	91	31	27	30	88
3	28	26	25	79	22	22	25	69
4	12	15	18	45	20	20	22	62
5	14	14	17	45	23	24	24	71
6	27	23	24	74	20	21	21	62
7	11	14	15	40	20	18	19	57
8	12	19	20	51	22	23	26	71
9	26	22	27	75	26	23	25	74
10	27	26	26	79	23	23	20	66
11	22	23	26	71	22	22	22	66
12	24	25	25	74	25	22	26	73
13	28	27	28	83	31	31	32	94
14	22	25	26	73	26	27	23	76
15	21	20	18	59	23	26	27	76

41

Question #7
Is there a correlation between sequential order and score?

- Results from top 80 candidates, 2025 Sergeant

	Rater 1 Total	Rater 2 Total	Rater 3 Total	Board 1 Total
Correlation with ID #	-0.32	-0.37	-0.47	-0.40
Probability	0.004	0.0007	0.00001	0.0002

- Board 1 (Raters 1 -3) scored in order of ID #

42

Question #7
Is there a correlation between sequential order and score?

- Results from top 80 candidates, 2025 Sergeant

	Rater 4 Total	Rater 5 Total	Rater 6 Total	Board 2 Total
Correlation with ID #	-0.42	-0.47	-0.28	-0.42
Probability	0.11	0.24	0.047	0.09

- Board 2 scored in reverse order of ID #

43

Question #7
Is there a correlation between sequential order and score?

- Yes.
- There is a negative correlation
- What to make of this?

44

Question #7
Is there a correlation between sequential order and score?

- It seems raters become more demanding over time
- Perhaps raters develop a clearer idea of the ideal answer as they grade more candidates

45

More Attempts to Understand

- Where to go from here?
- Suggestions for practice
- Further research

46

Guide for Practice

- **New Gold Standard**: Two Board Scoring
  - Replaces within board agreement
- Don't assume grades are valid or accurate because rater agreement is high
- Build in research component to BARS grading
  - Grading order, rater disagreement reasons

47

Suggestions for Research

- Post exam exploration of rater disagreements
- Seek collaboration with other testing groups
- Look at more job titles

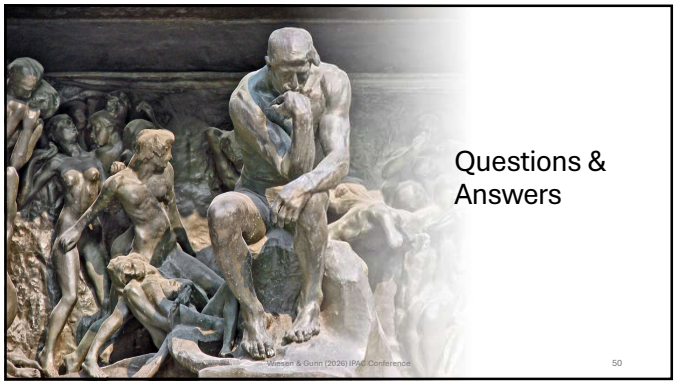
48

Powerpoints Online

https://ipacweb.org/library

https://appliedpersonnelresearch.com
/papers

49



Questions & Answers

50

References

- Jackson, D.J.R., Blair, M.D. & Ingold, P.V. (2024). Assessment centers: reflections, developments, and empirical insights. *Industrial and Organizational Psychology,* 17, 149-153.
- Jackson, D.J.R., Michaelides, G., Dewberry, C., Nelson, J. & Stephens, C. (2022). Reliability in assessment centres depends on general and exercise performance, but not on dimensions. *Journal of Occupational and Organizational Psychology,* 95, 739-757.
- Jackson, D.J.R., Michaelides, G., Dewberry, C. & Yang, W. (2024). The expert assessor perspective on assessment center taxonomies. *Human Performance*, Vol. 38, No. 1, 1-27.

51

References

- Wiesen, J. P. (2025). Reliability of BARS: New Data from Two-Board Grading. International Personnel Assessment Council Annual Conference, Atlanta, GA, United States.
- Wiesen, J. P. (2026, April 30). The BARS Inter-rater Reliability Illusion: a Within Versus Between-panel Comparison [Poster]. Society for Industrial and Organizational Psychology Annual Conference, New Orleans, LA, United States.

52