Reliability of BARS: New Data from Two-Board Grading

By
Joel P. Wiesen, Ph.D.
jw@jpwphd.com

2025 IPAC Annual Conference
Atlanta, GA, 7/28/2025 (updated 9/2/2026)

Who Am I?

- Independent consultant
- Worked 15 years for the MA civil service
- A scientist-practitioner
- 48 year member of IPAC
- 50 years working in the field of testing

Sequence of this Presentation

- New data on BARS reliability
- Attempts to understand the data
- History of the BARS approach
- Describe collection of these BARS data
- Suggest implications of the new data
- Solicit your ideas on where to go from here
- Some of my ideas on where to go

New Data on BARS Reliability

- BARS used to grade job exercises
- Data from 12 job exercises over 10 years
- High reliability within rater boards did not generally lead to high correlation between boards

Data from One Exercise

Reliability Within Board 1	This is the reliability for Board 1 for a certain exercise. It is based on the Spearman-Brown Prophecy formula (based on pairwise correlations for the 3 raters)
0.92	

Data from One Exercise

Reliability Within Board 1	Reliability Within Board 2	The second column shows the inter-rater reliability for the three raters in Board 2.
0.92	0.92	

Data from One Exercise

Reliability Within Board 1	Reliability Within Board 2	Expected Corr. Between Boards	The expected correlation between the Boards = the square root of the product of Board reliabilities
0.92	0.92	0.92	

Prediction Formula

$$r_{xy} = \sqrt{(r_{xx})(r_{yy})}$$

Source: Gregory (2011), page 114

Data from One Exercise

Reliability Within Board 1	Reliability Within Board 2	Expected Corr. Between Boards	Observed Corr. Between Boards	Observed correlation between Boards is lower
0.92	0.92	0.92	0.80	

Data from One Exercise

Reliability Within Board 1	Reliability Within Board 2	Expected Corr. Between Boards	Observed Corr. Between Boards	Percent (observed/ expected)	0.46/ 0.92 = 0.50
0.92	0.92	0.92	0.80	87%	

Data from One Exercise

Reliability Within Board 1	Reliability Within Board 2	Expected Corr. Between Boards	Observed Corr. Between Boards	Percent (observed/ expected)	Shrinkage
0.92	0.92	0.92	0.80	87%	13%

Pre and Post COVID-19

- Data do not all fit on one slide
- First slide will show pre-COVID-19
- Second slide will show post-COVID-19

Independent Boards

- Each exercise rated by two boards
- Each exam had two exercises
- Thus, four independent boards per exam
- Three raters on each board
- 12 raters for each person each year
- 2 boards x 2 exercises x 3 raters = 12

Reliability Within Board 1	Reliability Within Board 2	Expected Corr. Between Boards	Observed Corr. Between Boards	Percent (observed/ expected)	Shrink-age
0.92	0.92	0.92	**0.80**	87%	13%
0.89	0.86	0.87	**0.65**	74%	26%
0.89	0.95	0.92	**0.48**	52%	48%
0.89	0.87	0.88	**0.77**	87%	13%
0.88	0.80	0.84	**0.76**	90%	10%
0.85	0.89	0.86	**0.66**	76%	24%
0.97	0.91	0.94	**0.44**	47%	53%
0.97	0.94	0.96	**0.66**	69%	31%

Post COVID-19 Data

Reliability Within Board 1	Reliability Within Board 2	Expected Corr. Between Boards	Observed Corr. Between Boards	Percent (observed/ expected)	Shrink-age
0.86	0.90	0.88	**0.48**	55%	45%
0.96	0.96	0.96	**0.71**	74%	26%
0.94	0.94	0.94	**0.64**	68%	32%
0.94	0.94	0.94	**0.70**	74%	26%

Means

Reliability Within Board 1	Reliability Within Board 2	Expected Correlation Between Boards	Observed Corr. Between Boards	Percent (observed/ expected)	Shrink-age
0.91	**0.91**	0.91	**0.65**	0.71	0.29

Attempt to Understand the Data

- History of the BARS approach
- Collection of these BARS data
- Some implications of the new data
- Solicit your ideas on where to go from here
- Some of my ideas

History of the BARS Approach

- BARS were developed to address the low reliability of job performance ratings
  - Average reliability of .50 for supervisor evaluations (Conway & Huffcutt, 1997).
- BARS introduced to evaluate the job performance of nurses
  (Smith and Kendall, 1963)

Reliability Limits Validity

- Validity is always less than the square root of test reliability

$$r_{xy} = \sqrt{(r_{xx})(r_{yy})}$$

Anchor from Smith & Kendall

- If this nurse were admitting a patient who talks rapidly and continuously of her symptoms and past medical history, could be expected to look interested and listen.
- Better than: Good relationship with patients
- Better than: Good job performance

BARS Methodology

1. Identify dimensions
2. Define low, high, and acceptable perf.
3. Create list of potential anchors
4. Judges sort anchors into dimension
5. Other judges confirm or prune anchors
6. Judges rate potential anchors
   – Anchors with high variance are discarded

Red Flag in Original Article

- "The retranslation procedure, however, eliminated many items and several qualities ... **Communication Skills** ... [were] **eliminated** at this [retranslation] step." (Smith and Kendall, 1963, pg 152)
- But patient teaching is central to nursing

Red Flag in Original Article

- "Communication Skills items were allocated to Skill in Human Relationships or Conscientiousness because the items involved, to a large extent, either explaining to patients or keeping records." (Smith and Kendall, 1963, pg 152)
- Dimension multi-faceted

1970s Surge of Rating Methods

- Modest support for BARS
- "there is some evidence to suggest that behavioral anchors are better than numerical or adjectival ones" (Landy & Farr, 1980, Page 88, col 1)
- "the BARS method has not been supported empirically." (Page 85)

Reliability of BARS

- BARS are more reliable than other rating methods, but **only marginally**, with reliability of .77 vs .73
  (Levashina et al., 2014, page 273)
- Some find higher BARS reliability: 0.84
  (McDaniel et al., 1994, Table 2 for structured interviews)
- Some find lower non-BARS reliability: 0.46
  (Dunnett & Montiwidlo, 1976, Table 4)

Reliability of BARS

- BARS more subjective than checklists
  – (Whetzel, Rotenberry & McDaniel, 2014)

Why Are BARS Popular?

- Testing driven by legal concerns
- BARS more precise than generic anchors
  – Unacceptable
  – Satisfactory
  – Good
  – Excellent
- Little or no adverse impact
  (Buckley et al., 2004; Levashina et al., 2014)

Collection of These BARS Data

- Police promotional exams (biannual)
- Practical exercises presented orally
- Two independent rater boards
- Initial and final ratings made
- 9 point BARS

Collection of These BARS Data

- Four 9 point BARS scales
- Typical dimensions rated:
  - Oral Communication
  - Command Presence
  - Interpersonal Relations
  - Supervision
  - Problem Analysis
  - Problem Resolution

Collection of These BARS Data

- Raters train together with role playing
  - Number of candidates varied, around 100
- Initial grades without discussion
- Discuss differences in ratings of 4+ points
- Final grades do not require agreement
- Pre-COVID-19 live and video grading
- Post-COVID-19 all boards do video grading

Collection of These BARS Data

- Two exercises in each exam
- Different raters on the two boards
- Four BARS per exercise
- The 4 BARS ratings are averaged

One Additional Analysis

- Oral communication ratings showed the same size disagreements between pairs of raters within a board as other dimensions

Implications of the New Data

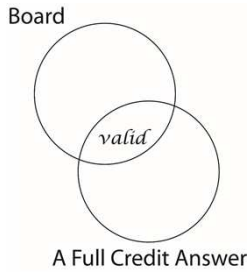
- Each Board agreed internally on the grading
- The 2 Boards used different grading criteria
  - Overlap in grading criteria was too low

Classic Test Theory

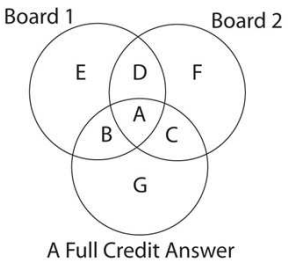
$X = T + E$

X = observed score
T = true score
E = error

Venn Diagram of Validity



Expanded Venn Diagram



Implications of the New Data

- High within-board reliability is still good
  - Not as good as we once thought!
- High reliability may not indicate validity
- Low between-board reliability indicates low validity
- Using two boards probably better than one

Reliability Limits Validity

- Validity is always less than the square root of test reliability

$$r_{xy} = \sqrt{(r_{xx})(r_{yy})}$$

Possible Source of Disagreement

- Major unanswered question about BARS
- Observe anchors at several scale points
- What rating to give?

Possible Source of Disagreement

- BARS are not a Guttman scale
- Definition of Guttman scale
  - Choosing one scale point implies satisfying all lower points
  - Can run a 4 min. mile implies answer to 5 min

(Guttman, 1944).

Non-Guttman BARS

1. Exhibits reluctance to make a decision even when decision is appropriate.
3. Gathers information before making decisions but may overlook relevant data.
5. Gathers all needed information before making decision.
7. Identifies and chooses appropriate solutions so that problem is eliminated or alleviated.

Guttman BARS Facets

- Identify facet KSAPs
- Create Guttman anchors for each KSAP
- Identify highest score for each KSAP
- Provide guidance on final rating
  - Need high score on both facets for high score
  - Averaging facet scores usally NOT appropiate
  - Seriously low score on a KSAP drives overall

Why Low Inter-Board Reliability

- Your ideas
  - How to understand data
  - How to improve grading
- Some of my thoughts

Your Ideas

- Floor is open

Why Low Inter-Board Reliability

- Is inter-board reliability a function of
- Question content
- BARS dimension
- BARS scale anchors

Why Low Inter Board Reliability

- Why were 2 of the between panel correlations well predicted but the other 10 showed shrinkage?
- What caused the difference in grading criteria between panels given that all the panelists were trained together.

Why Low Inter Board Reliability

- Was the shrinkage the same for the 4 BARS as for the average.

Why Low Inter Board Reliability

- Can raters explain differences
- Can monitors see between board differences

Conclusions

- BARS are used widely and are considered the best approach in many situations
- High inter-rater reliability is taken to signal accuracy of the BARS grading
- New data show that high inter-rater reliability within a board may not translate to high correlation between boards

Recommendations

- Use two rater boards when using BARS

In Closing

- Improve rating of job exercises
- Find reasons for low between-board correlation
- Refine current approaches to rating
- Invent new approaches to rating

Contact Information

- Session URL: **http://jpwphd.com/ipac2025**
- Email: **jw@jpwphd.com**
- Telephone: **(617) 244-8859 (land/no text)**
- **Email and telephone calls welcome!**
- **Q&A's**

IPAC Member Forum

- Place to discuss discretely
  - Open to members only
- Access it on **ipacweb.org**
  - click on ***three dots*** at right of menu bar
  - select ***Members Only***
  - select ***IPAC Member Forum***

Learning Objectives

- Describe the major reason the that the BARS approach was originally proposed.
- Describe why the use of two rating boards has challenged accepted thinking on the reliability of BARS grading.
- Describe at least one research approach that could be used to identify reasons for lower reliability between rater panels vs within

Q&A

Presentation Citation

Wiesen, J. P. (2025). Avoiding Pitfalls: Identifying and Addressing Flawed, Implicit Assumptions in Testing. *International Personnel Assessment Council*, Atlanta, GA, United States.

References

- Buckley, M. R, Jackson, K. A., Mark C. Bolino, M. C., Veres, III, J. G. & Feild, H. S. (2007). The Influence of Relational Demography On Panel Interview Ratings: A Field Experiment. *Personnel Psychology, 60,* 627–646.

References

- Dunnette, M. D., & Motowidlo, S. J. (1976, November). *Police selection and career assessment* (Report No. ED 130 042 / CE 008 247). Personnel Decisions, Inc. https://eric.ed.gov/?id=ED130042

References

- Gregory, R.J. (2011) *Psychological Testing History, Principles, and Applications (6th ed.)* Boston, MA: Allyn & Bacon.

References

- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150. https://www.jstor.org/stable/2086306

References

- Landy, F. J. & Farr, J. L. (1980). Performance Rating. *Psychological Bulletin, 87*, 72-107.

References

- Levashina, J., Hartwell, C. J., Morgeson, F. P. & Campion, M. A. (2014) The structured employment interview: narrative and quantitative review of the research literature. *Personnel Psychology*, *67*, 241–293.

References

- McDaniel, M. A., Whetzel, D. A., Schmidt, F. L. & Maurer, S. D. (1994). The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis. *Journal of Applied Psychology, 79*, 599-616.

References

- Smith, P. C. & Kendall, L. M. (1963) Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155.

References

- Whetzel, D. L., Rotenberry. P. F. & McDaniel, M. A. (2014). In-basket Validity: A systematic review. *International Journal of Selection and Assessment, 22*, 62-79.