The BARS Inter-rater Reliability Illusion: a Within Versus Between-panel Comparison

SHORTENED TITLE
The BARS Inter-Rater Reliability Illusion

ABSTRACT

We misinterpret the high inter-rater reliability often obtained with Behaviorally Anchored Rating Scales (BARS) when we assume that it indicates that scoring is reliable and, by inference, accurate. New data indicates that high inter-rater reliability can be unique to one grading panel. Data from the scoring of 12 police promotional examination exercises show that high within-panel reliability sometimes coexists with low between-panel grade correlations. It may be that individual rater panels develop idiosyncratic grading criteria, despite all panelists receiving the same training.

WORD COUNT
1,193

The BARS Inter-rater Reliability Illusion: a Within Versus Between-panel Comparison

Inter-rater reliability of grading with a BARS is generally high (e.g., 0.77, Levashina et al., 2014, page 273; and 0.84, McDaniel et al., 1994, Table 2 for structured interviews). Low inter-rater reliability is taken as an indication of grading deficiency and high inter-rater reliability is considered desirable (preferaby 0.95 or higher, and at least 0.90, according to Nunnally & Bernstein, 1994, page 265). However, there is no systematic literature that evaluates BARS rater-reliability across two rater panels scoring the same candidates on the same exercises. I encountered such data in my consulting work, as reported below.`

Method

Data were available for 12 exercises administered as part of promotional examinations for police sergeant that were given biannually from 2013 to 2023. Each examination year included an oral panel component consisting of two new, independent exercises. Each exercise was graded by two independent rater panels, each composed of three raters. Different raters graded the two exercises. The raters were all current Sergeants or Lieutenants in comparable police departments located in various states other than the location of the examination. A small number of raters graded more than one examination. Some candidates took the promotional examination more than once, in different years. The average number of candidates was 95 (see Table 1).

The exercises simulated the work of a police sergeant including, for example, a one-on-one discussion with a subordinate experiencing some difficulties, addressing a group of subordinates on a critique of an emergency scene, or addressing a civilian or a group of civilians on topics such as community policing or crime in the neighborhood.

Each exercise was graded on four dimensions. The dimensions rated varied by exercise and year and typically included such dimenstions as: Oral Communication, Interpersonal Relations, Supervision, Problem Analysis, and Problem Resolution. A candidate's score for an exercise was the average of 12 BARS ratings (3 raters on 4 nine-point BARS scales).

Rater training was conducted before any rating of actual test takers. This training included: job duties of a Sergeant, the BARS methodology, focusing on behaviors rather than general impressions, common rater errors, and practice in grading the exercises. In live rating, initial ratings were made prior to any discussion. Large discrepancies among the raters on any dimension were discussed, followed by the raters making their final ratings. Consensus was not required. Only the final ratings are reported below but the pattern is very similar for the initial ratings.

Before COVID, from 2013 to 2019, one panel both administered and graded the exercise and the second panel graded the candidates based on video recordings. After COVID, in 2021

and 2023, the exercises were administered by video and both panels graded the candidates based on the video recording.

The Spearman-Brown reliability for the pairs of panels was used to predict the expected correlation between-panels using this formula (Ghiselli, Campbell, & Zedeck, 1981, page 242):

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

Results

Within-panel Spearman-Brown reliability was high was high, averaging 0.91. The predicted correlation between-panels, based on within-panel relibility, was also high, averaging 0.91. However, the observed correlation between-panels was lower, averaging 0.65; never above 0.80 and below 0.50 for one-fourth of the exercises. The ratio of the observed to the expected between-panel correlations, expressed as a percent, averaged 71%, meaning there was a 29% average shrinkage from the expected to the observed between-panel correlation. (See Table 1.)

Table 1. Within-panel Reliability and Between-panel Correlation by Year and Exercise								
Year	Exercise	N	Reliability Within Panel 1	Reliability Within Panel 2	Observed Correlation Between Panels	Expected Correlation Between Panels	Percent (observed/ expected)	Shrinkage
2013	1	86	0.92	0.92	0.80	0.92	87%	13%
2013	2	86	0.89	0.86	0.65	0.87	74%	26%
2015	1	109	0.89	0.95	0.48	0.92	52%	48%
2015	2	109	0.89	0.87	0.77	0.88	87%	13%
2017	1	66	0.88	0.80	0.76	0.84	90%	10%
2017	2	66	0.84	0.88	0.66	0.86	77%	23%
2019	1	124	0.97	0.91	0.44	0.94	47%	53%
2019	2	124	0.97	0.94	0.66	0.96	69%	31%
2021	1	104	0.86	0.90	0.48	0.88	55%	45%
2021	2	104	0.96	0.96	0.71	0.96	74%	26%
2023	1	80*	0.94	0.94	0.64	0.94	68%	32%
2023	2	80*	0.94	0.94	0.70	0.94	74%	26%
Means		95	0.91	0.91	0.65	0.91	71%	29%

*Only 80 of the 107 candidates were rated; those with the highest scores on the other three examination components

Discussion

Possible explanations for these findings can be visualized with a Venn diagram (see Figure 1). On the positive side, the use of two rater panels may increase validity (see areas B and C) as each panel may capture valid aspects missed by the other panel. On the negative side, the observed agreement between panels may include invalid variance (see area D), due to scoring of invalid factors (e.g., a person's physical size). Areas E and F must include systematic as well as random error since the within-panel reliability is always high. The systematic error not shared between panels may reflect idiosyncratic, panel-specific grading criteria (perhaps "similar to me" error).

The lower circle in Figure 1 is labeled "A full credit answer" rather than "The correct answer" because there may be more than one correct way to address complex work situations. The possibility of multiple correct answers could explain some of the shrinkage seen in Table 1, but is not considered further here.

One implication of this research is clear: using two BARS rating panels to grade complex job simulation exercises is now highly desirable, or even best practice.

Future research should try to identify the cause of low between-panel correlation seen in the grading of some exercises. Such future research might include:
(1) having each pair of panels meet together to talk about the candidates whose scores show the most discrepancies,
(2) having the agency staff who facilitated each pair of panels meet to talk about the candidates whose scores show the most discrepancies,
(3) having the agency staff who facilitate the rating panels swap places midway through the grading to try to detect differences in grading criteria between panels,
(4) doing a content analysis of the various exercises to try to identify the type of exercise that have low or high between-panel correlations, and
(5) analyzing the data to try to identify the dimensions associated with low or high between-panel correlations.

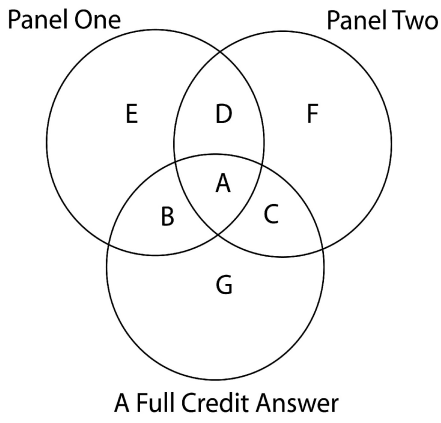

Figure 1. Venn Diagram Showing Possible Valid and Invalid Variance in Grading

Key:
A: Valid variance identified by both rater panels
B. Valid variance identified by panel 1 but not panel 2
C. Valid variance identified by panel 2 but not panel 1
D. Invalid agreement between the two panels
E. Invalid variance exhibited only by panel 1 (both random and panel-specific)
F. Invalid variance exhibited only by panel 2 (both random and panel-specific)
G. Valid variance missed by both panels

Conflict of Interest Statement

I gained access to the examination statistics in the first 5 columns of Table 1 in my consulting role. These statistics were computed by the client as part of their standard post-examination analyses. The client did not influence my interpretation or the content of this article. I have no other potential conflicts of interest to disclose.

References

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: Freeman.

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.

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.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.) New York: McGraw-Hill.