First Steps in Validating Explicit and Implicit On-the-Job Performance Assessments

DR. LILLIAN K. E. ASIALA &
DR. JAMES E. MCCARTHY

Distribution Statement A: Cleared for public release. Case # AFRL-2022-2668

Acknowledgements

This research was funded by Government contract: FA8650-18-C-6932, and summarizes work that Sonalysts, Inc. performed for KBR.

The views expressed in this presentation are mine and do not necessarily reflect the official policy or position of the Department of the Air Force, the Department of Defense, or the U.S. government.

Sonalysts Team Members for this effort

- Dr. James McCarthy
- Dr. Lillian Asiala
- Dr. Teena Garrison
- Dr. Melinda McGurer

Overview

Problem Space: Human performance measurement in uncontrolled environments.

Measurement Context: Measurement with the goal of improved career alignment in the U.S. Air Force.

Question: Which measurement approach is best (explicit v. implicit)?

Data Collection/Performance Evaluation

Data Analyses

Conclusion

Measurement in the “Wild”

Measuring performance is most authentic in the “real world.”

What is the best approach to measurement?

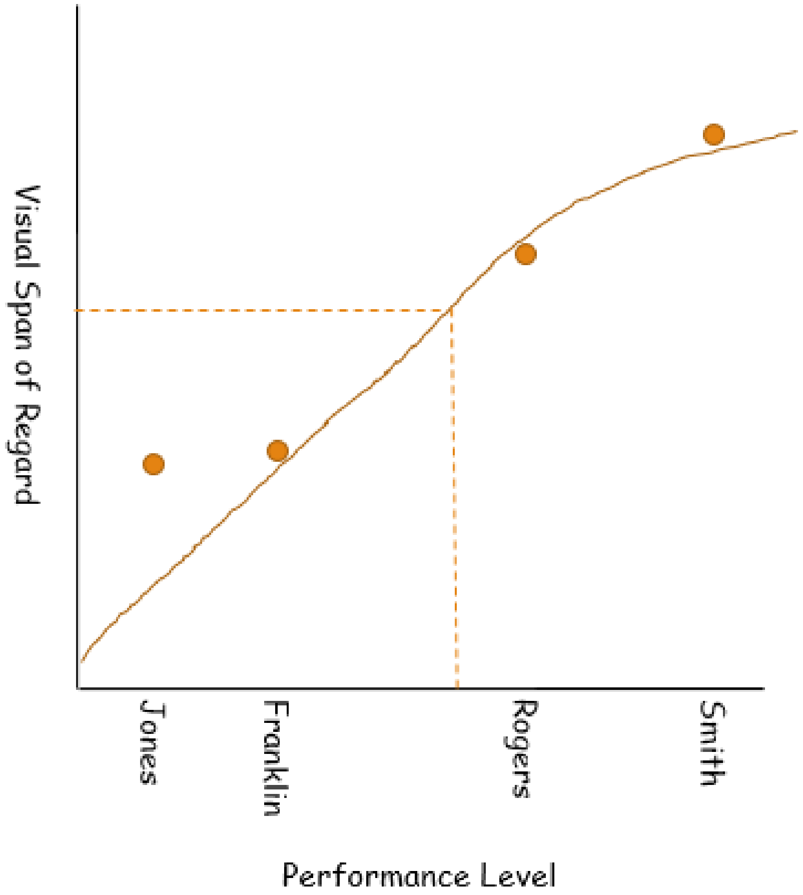
- Quantitative measures of specific actions?
- Qualitative “gut feeling” judgements from experts?
- Something in-between?

Goal: Validate three levels of flexible, process-oriented measures of performance.

PBAO

Measurement validation context: PBAO

- Goal: Use GSPT to identify attribute levels that predict operational performance for U.S. Air Force careers.
- Sub-goal: Evaluate operator performance in a realistic mission environment.



Performance Measurement Tool Qualities

Reliability: One performance is given similar scores.

Validity: Match between quality of performance and score.

Discriminability: Performances can be differentiated from one another.

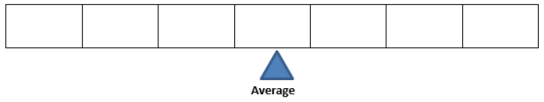
Explicitness: Specifying processes in the performance environment.

Measurement Explicitness

Implicit

Explicit

Graphic Rating Scale (GRS)

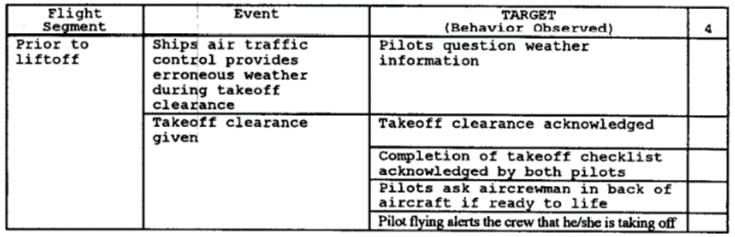


Behaviorally Anchored Rating Scale (BARS)

Better Performance
Worse Performance

Score	Sample Behaviors
7	Specific landmark behavior. Specific landmark behavior. Specific landmark behavior.
6	Behaviors similar to those in 5 and 7.
5	Specific landmark behavior. Specific landmark behavior. Specific landmark behavior.
4	Behaviors similar to those in 5 and 3.
3	Specific landmark behavior. Specific landmark behavior. Specific landmark behavior.
2	Behaviors similar to those in 1 and 3.
1	Specific landmark behavior. Specific landmark behavior. Specific landmark behavior.

Performance Checklist



Flight Segment	Event	TARGET (Behavior Observed)	4
Prior to liftoff	Ships air traffic control provides erroneous weather during takeoff clearance	Pilots question weather information	
	Takeoff clearance given	Takeoff clearance acknowledged	
		Completion of takeoff checklist acknowledged by both pilots	
		Pilots ask aircrewman in back of aircraft if ready to life	
		Pilot flying alerts the crew that he/she is taking off	

From Fowlkes *et al*. (1998)

Data Collection: Exercise Recordings

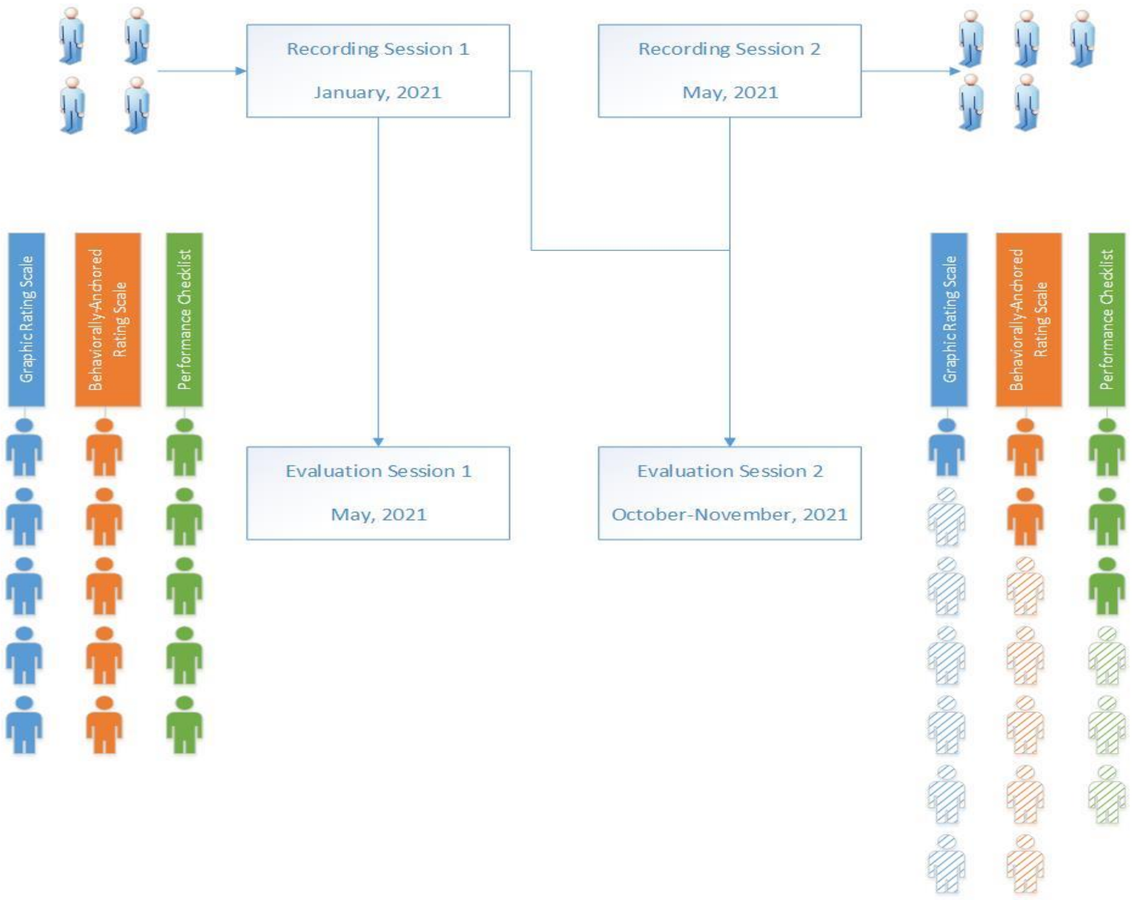
Training exercises conducted in high fidelity simulation.

Post training exercise survey:

- Role
- Performance Quality
- Difficulty

Nine participants recorded over two data collection sessions.

Data Collection



Recording Session 1

January, 2021

Recording Session 2

May, 2021

Evaluation Session 1

May, 2021

Evaluation Session 2

October-November, 2021

Analyses

Construct Validity

Interrater Reliability

Face Validity

Construct Validity

Compares new measures to:

- Existing measures
- Future outcomes (predictive validity)

No comparable existing measures.

Pearson correlations calculated for new measures.

	BARS	Checklist
GRS	.637, *p* = .065	.822, *p* = .007
BARS		.644, *p* = .061

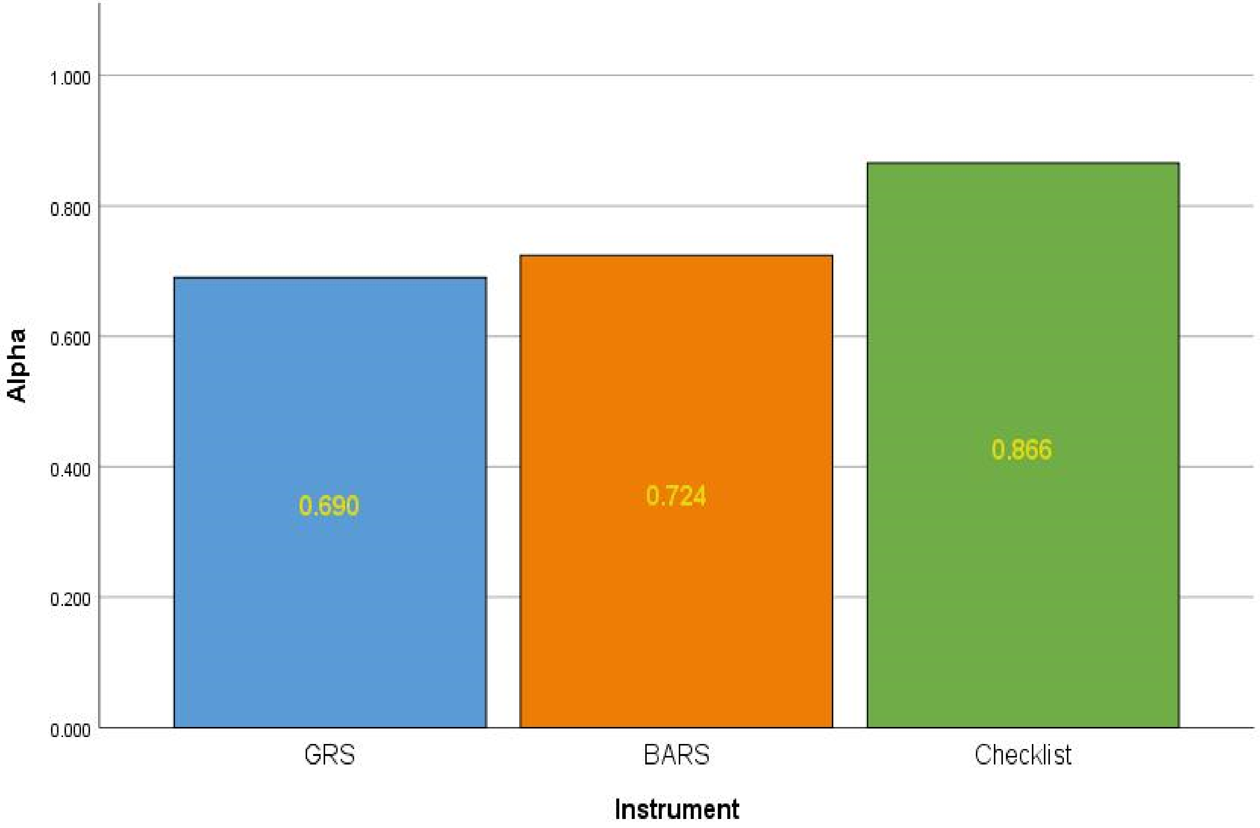
Distribution Statement A: Cleared for public release. Case # AFRL-2022-2668

Interrater Reliability

Cronbach's Alpha

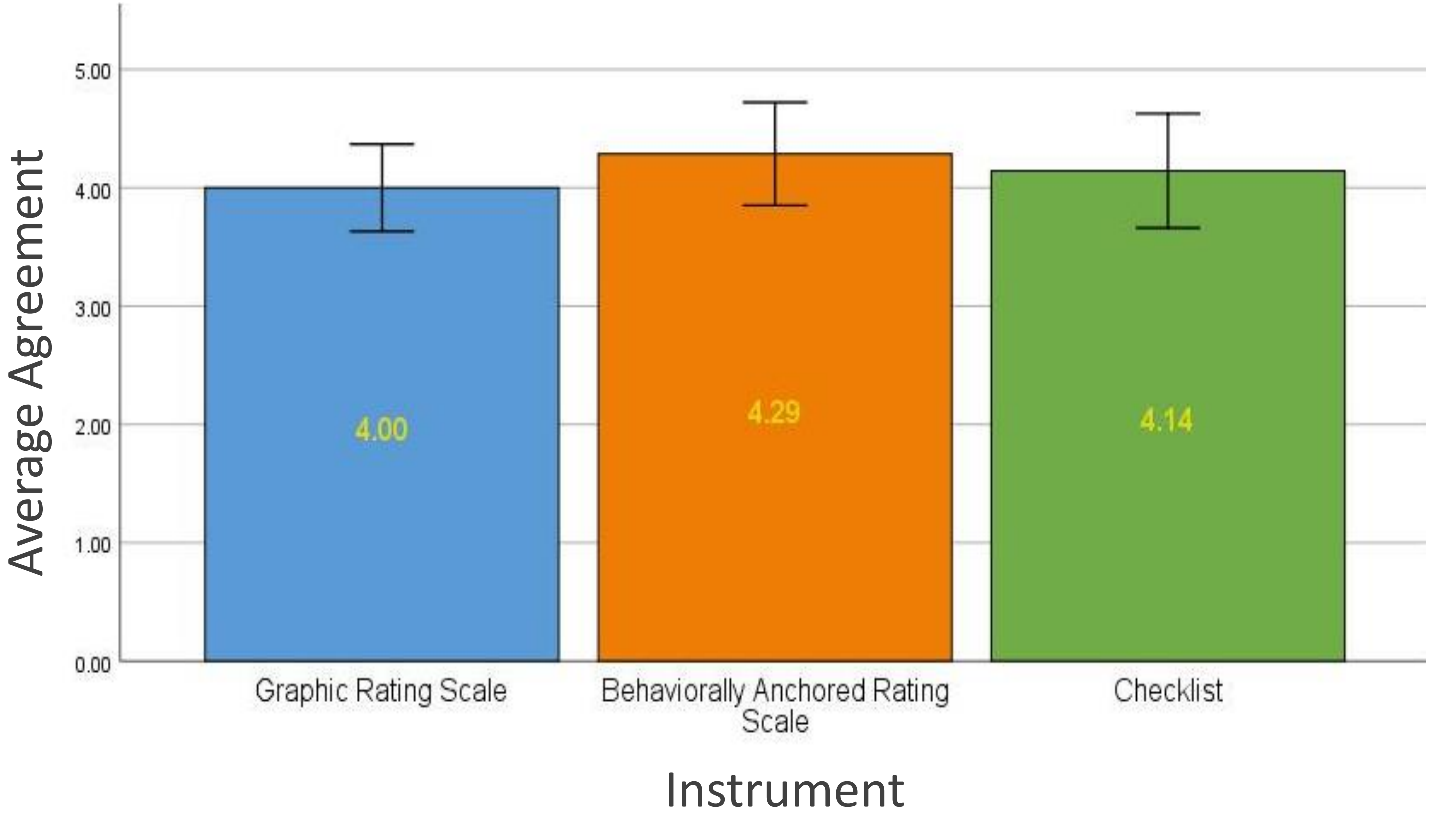
- Measure of internal consistency.
- Calculated for evaluators.
- Averaged for each measure.

As measures becomes more explicit, Cronbach's Alpha becomes larger.



Face Validity

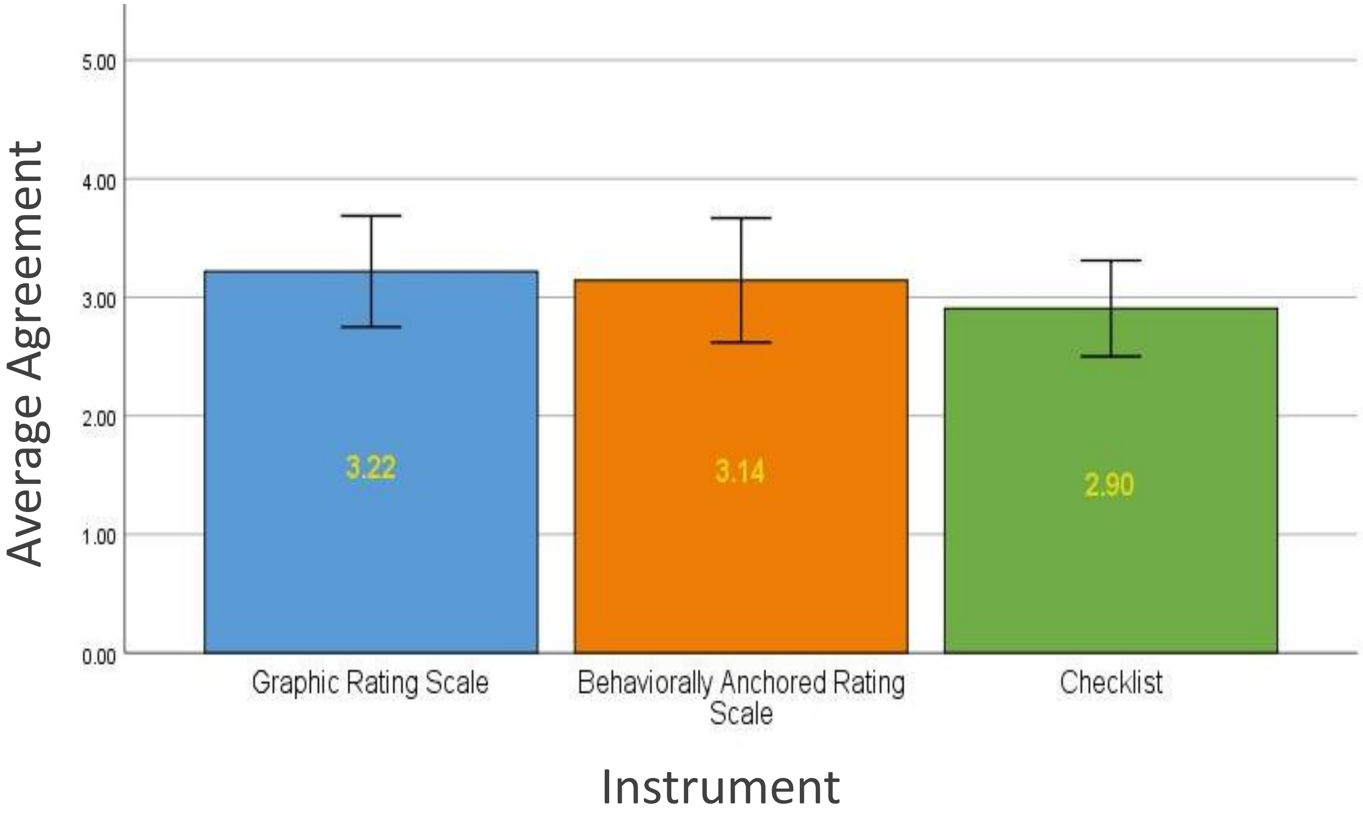
“I think that experts can use this measure to accurately rate the performance of operators in a variety of missions.”



Error Bars: 95% CI

Face Validity

“If different experts used this measure to rate the performance of an operator completing a given mission, they would produce very different scores.”



Error Bars: 95% CI

Conclusions

All measures were positively correlated, with the highest correlation between GRS and checklist scores ($r(7)=.822$, *p = .007).*

Members of the operational community generally approved of the measures (though confidence in application was lukewarm).

Criteria for GRS scores differs from the prioritization of behavioral indicators on the checklist.

The checklist and GRS have different strengths.

- The GRS may have a better "return on investment" considering development cost.
- The checklist is more sensitive to observable aspects of performance (which could offer insights for training).

Additional research is needed to validate the trending relationships presented here.