(Conference Version)

BARS Rater-Reliability: Levels and Implications

Joel P. Wiesen, Ph.D.

Contact: **jpw@jpwphd.com**

2021 Annual IPAC Conference
Presented Virtually; 7/26/2021

Thanks to Michael Blair

- Prepared an introduction to BARS
- Could not present due to a conflict
- Provided valuable feedback to other presenters

Print and Audio Links

- PowerPoints (pre-conference to be posted)
- Audio recording (to be posted)
- **http://jpwphd.com/ipac2021**

Questions

- Please type questions in chat
- We will compile the chat questions and try to answer them at the end.

Outline of This Presentation

1. Three foci drove the origin of BARS
2. A few illustrative BARS in the literature
3. Level of inter-rater reliability
4. Stability of appointments based on BARS
5. Caution on statistical equating
6. Rater training
7. Why are BARS widely used?

1. Three Original Foci of BARS

- Clarify dimensions using behaviors
- Clarify scale values in terms of behaviors
- Involve 2 groups of SMEs in so doing

(Smith & Kendall, 1963, page 149)

Three Concepts Behind BARS

- Retranslation
- Rater agreement on scaling of anchors
- Wide applicability

Retranslation in Languages

- To translate an exam:
- One team translates into the new language
- Second team translates back
- Compare the retranslation and original
- One team cannot do both steps

Retranslation in BARS

- One team writes anchors for the dimensions
- Second team assigns anchors to dimensions
- One team cannot do both steps
- Retranslation should clarify dimensions and improve discriminant validity

Agreement on Scaling of Anchors

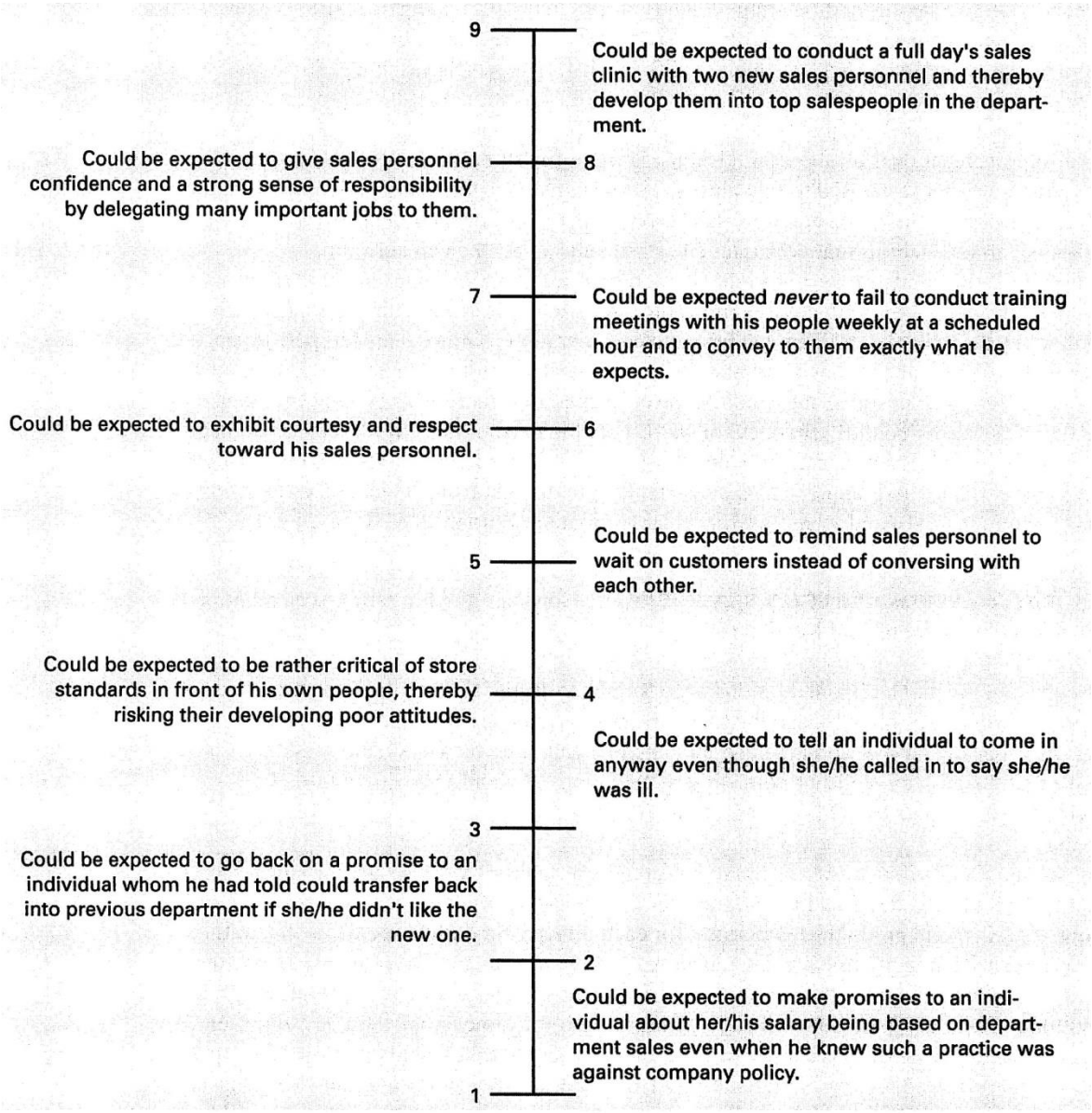
- Potential anchors rated by SMEs
- Selected anchors must have small variance
- Anchors might not improve reliability, since reliability is measured by correlation
- Anchors might improve accuracy
  - Often rater accuracy is not investigated

Wide Applicability

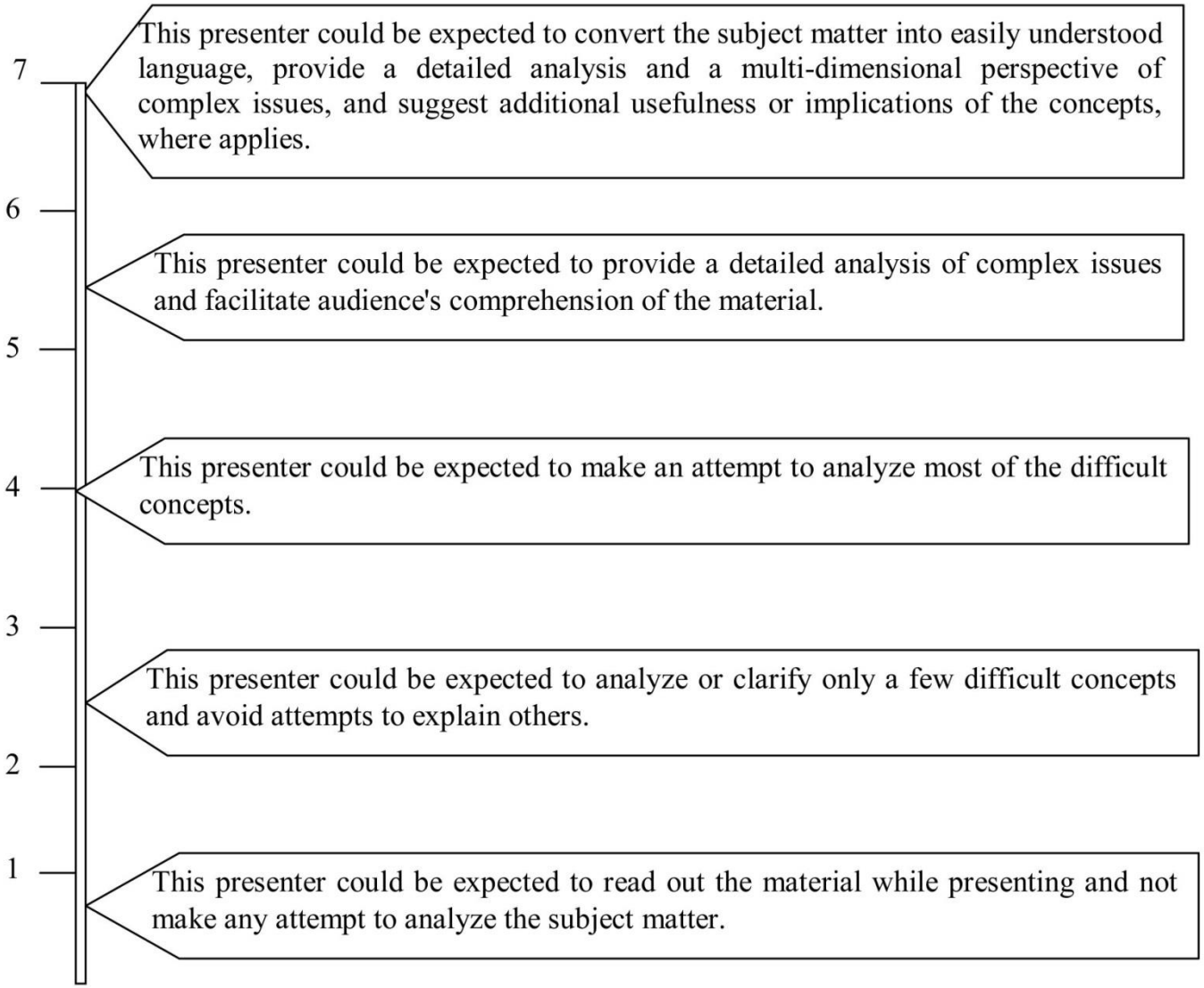
- Use “behavioral expectations” as anchors
- Avoid situation-specific anchors
- Goal: BARS that can be used repeatedly

2. A Few Illustrative BARS

- Many anchors vs few
- Descriptive categories vs none
  - Effective
  - Minimally passing



(Cascio & Aguinis, 2011, page 94)



(Debnath, Lee & Tandon, 2015, page 18)

Judgment - Observation and assessment of the situation and taking appropriate action.

The descriptions to the right are examples of behavior of individual patrol officers who are usually rated "High" on "Judgment" by supervisors.

High

Calls for assistance and clears the area of bystanders before confronting a barricaded, heavily-armed suspect.

Notices potentially dangerous situations before anything actually occurs.

Radios in his position and discontinues a high-speed chase before entering areas of high vehicle and pedestrian traffic, such as school areas.

The descriptions to the right are examples of behavior of individual patrol officers who are usually rated "Average" on "Judgment" by supervisors.

Average

Issues warnings instead of tickets for traffic violations which occur at particularly confusing intersections for motorists.

Permits traffic violators to explain why they violated the law and then decides whether or not to issue a citation.

Does not leave a mother and daughter in the middle of a fight just because no law is being violated.

The descriptions to the right are examples of behavior of individual patrol officers who are usually rated "Low" on "Judgment" by supervisors.

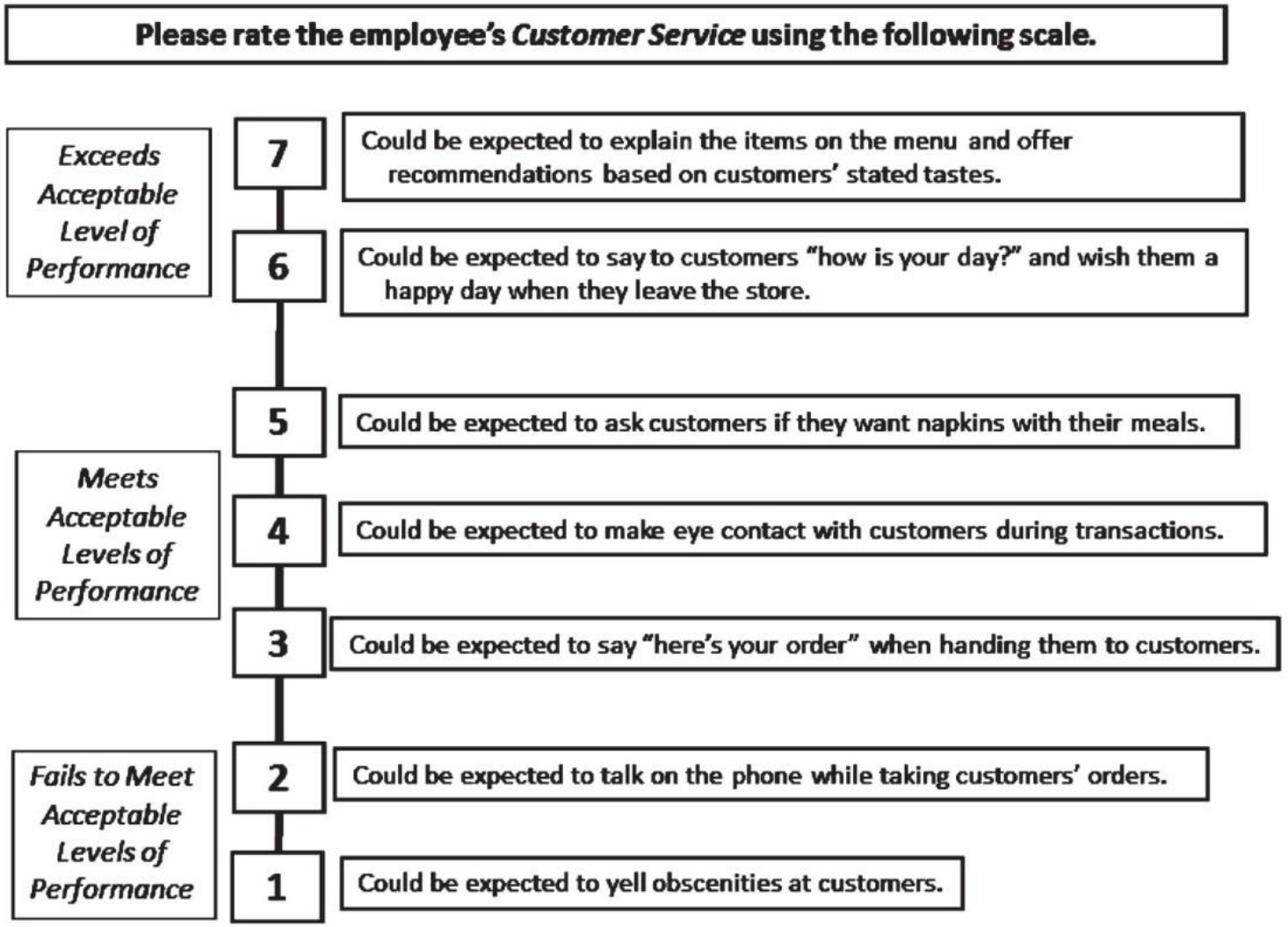
Low

Enters a building with a broken door window instead of guarding the exits and calling for a backup unit.

Does nothing in response to a complaint about a woman cursing loudly in a restaurant.

Continues to write a traffic violation when he hears a report of a nearby robbery in progress.

(Landy, 1977)

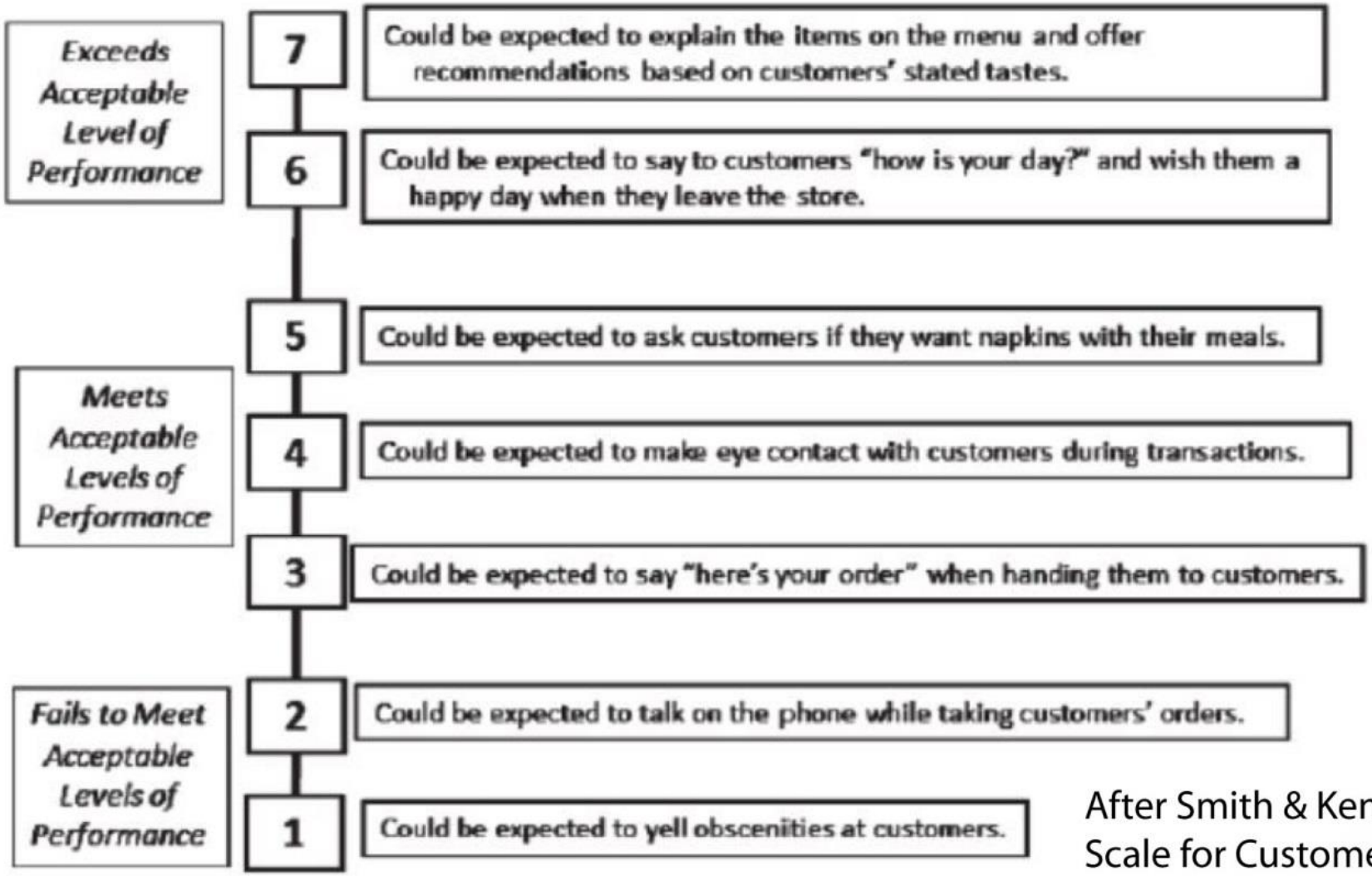


(Kell, et al., 2017)

Comprehensive Assessment of Team Member Effectiveness—Behaviorally Anchored Rating Scale (BARS) Version

	Your name					← Write the names of the people on your team including your own name. This self and peer evaluation asks about how you and each of your teammates contributed to the team during the time period you are evaluating. For each way of contributing, please read the behaviors that describe a "1", "3," and "5" rating. Then confidentially rate yourself and your teammates.
Interacting with Teammates	5	5	5	5	5	• Asks for and shows an interest in teammates' ideas and contributions. • Improves communication among teammates. Provides encouragement or enthusiasm to the team. • Asks teammates for feedback and uses their suggestions to improve.
	4	4	4	4	4	Demonstrates behaviors described in both 3 and 5.
	3	3	3	3	3	• Listens to teammates and respects their contributions. • Communicates clearly. Shares information with teammates. Participates fully in team activities. • Respects and responds to feedback from teammates.
	2	2	2	2	2	Demonstrates behaviors described in both 1 and 3.
	1	1	1	1	1	• Interrupts, ignores, bosses, or makes fun of teammates. • Takes actions that affect teammates without their input. Does not share information. • Complains, makes excuses, or does not interact with teammates. Accepts no help or advice.
Keeping the Team on Track	5	5	5	5	5	• Watches conditions affecting the team and monitors the team's progress. • Makes sure that teammates are making appropriate progress. • Gives teammates specific, timely, and constructive feedback.
	4	4	4	4	4	Demonstrates behaviors described in both 3 and 5.
	3	3	3	3	3	• Notices changes that influence the team's success. • Knows what everyone on the team should be doing and notices problems. • Alerts teammates or suggests solutions when the team's success is threatened.
	2	2	2	2	2	Demonstrates behaviors described in both 1 and 3.
	1	1	1	1	1	• Is unaware of whether the team is meeting its goals. • Does not pay attention to teammates' progress. • Avoids discussing team problems, even when they are obvious.

(Ohland, et al, 2012)



After Smith & Kendall, 1963
Scale for Customer Service

Planning, Preparing and Organizing Work: Plans work prior to performing tasks; reads blueprints carefully before beginning work; and anticipates tool and material needs of journeymen before being asked to order and retrieve materials.

7
6 This apprentice suggests that workers turn in material lists a day in advance so that the apprentice can have the material ready and organized for the following day.

5 This apprentice orders and retrieves materials from the truck before being asked.

4 This apprentice stacks fixtures, takes apart fixtures, and puts each fixture in a designated area for easy removal.

3
2 This apprentice uses shallow boxes for piping schoolrooms for lighting because they are available, rather than ordering appropriate material for the job.

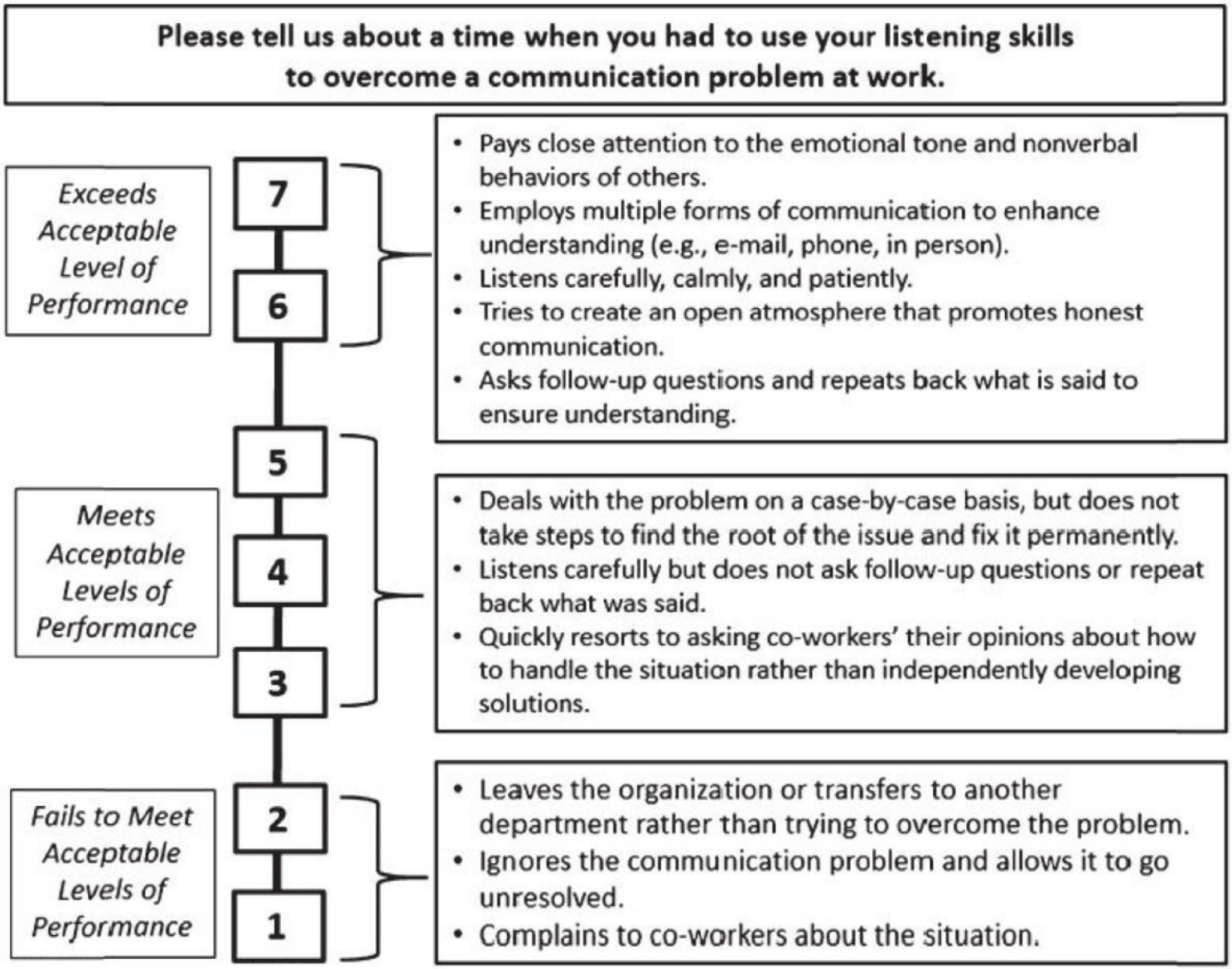
1 This apprentice cuts holes in walls without reviewing blueprints, resulting in major property damage.

Pulakos (1997), Fig 11.2

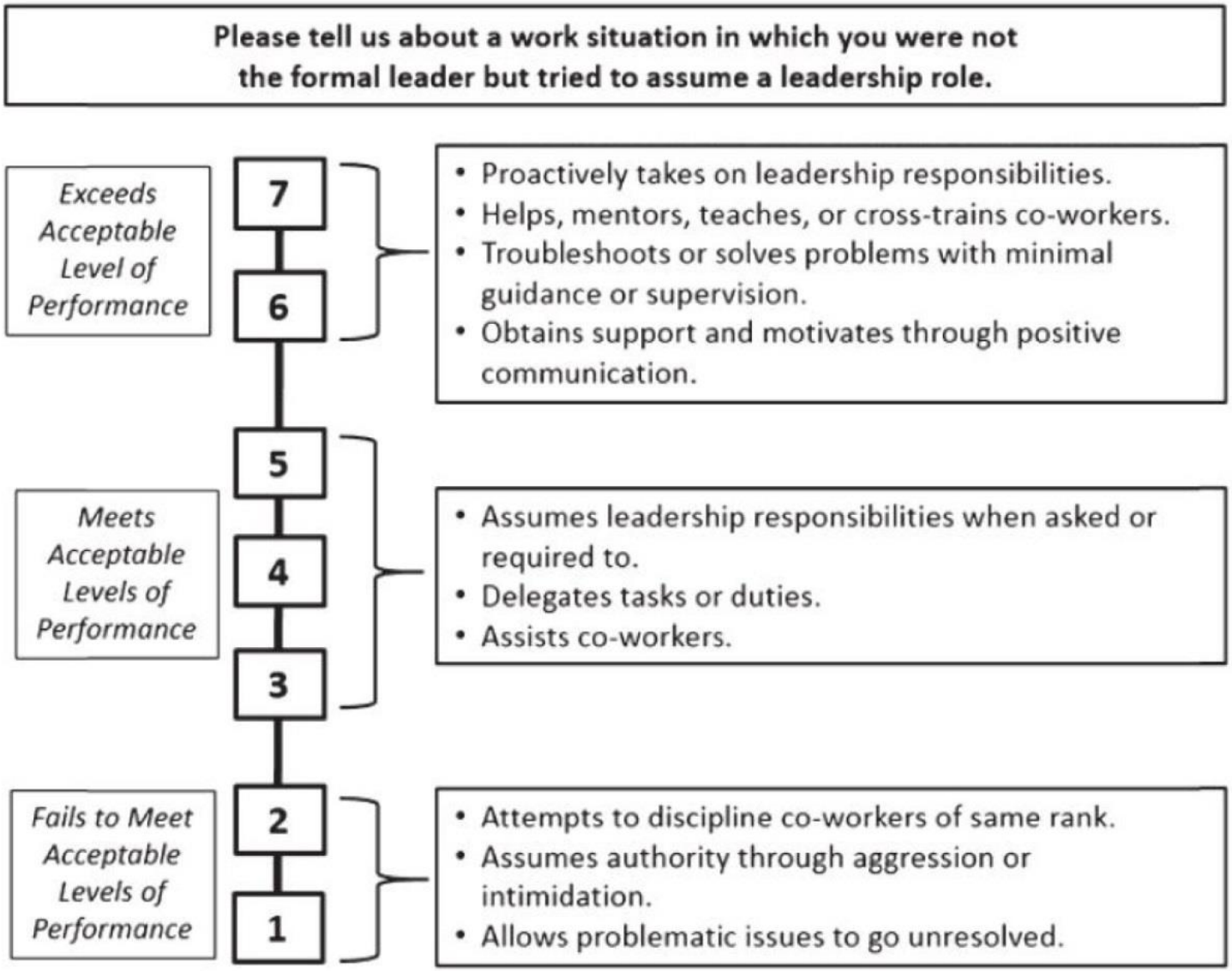
Please rate the apprentice on the following scale by reading the description of each performance level and selecting the number that most closely corresponds to the behavior exhibited by the ratee.

LOW	MEDIUM	HIGH
Starts to perform tasks without checking blueprints/plans; uses inappropriate but available materials for jobs; hurries through work before considering additional tasks that need to be performed before job is completed; misjudges time to complete tasks.	Plans tasks before performing them; organizes tools and materials so that they are easily retrievable when needed; writes down information needed for jobs so that work is completed efficiently; makes adjustments to material before starting jobs so that work is completed efficiently.	Reviews blueprints/plans before starting tasks; makes suggestions about organizing material so that preparation time is minimized; creates lists of tasks to be completed and orders material on own initiative; keeps foreman informed of progress of jobs; anticipates needs for jobs and retrieves tools and materials before they are requested.
1 2	3 4 5	6 7

Pulakos (1997), Fig 11.3



Kell, et al. (2017) Appendix B



Kell, et al. (2017) Appendix B

BARS for Layoff Decisions

- **Follows Processes and Policies** refers to how well the employee performs work in accordance with institutional policies and technical processes. … Below are some examples for you to consider when making your rating. You may consider this person's **Follows Processes and Policies** for other job tasks as well.
- (Bobrow, 2021, personal communication)

Sometimes follows company internal policies and procedures (e.g., HR, expenses, procurement, etc.), but has frequently been cited for taking short-cuts or not completely follow them.	Usually follows company institutional policies, procedures, and processes but occasionally takes short-cuts or does not completely follow them.	Consistently follows company institutional policies and procedures, procedures, and processes (company "Rules").
Consistently holds self and others to the highest standards of integrity in peer reviews, scientific evaluations, and recommendations, but has been cited for not doing so.	Usually holds self and others to the highest standards of integrity in peer reviews, scientific evaluations, and recommendations.	Consistently holds self and others to the highest standards of integrity in peer reviews, scientific evaluations, and recommendations.
①	③	⑤

(Bobrow, 2021, personal communication)

BARS for Layoff Decisions

- **Interpersonal Effectiveness** refers to how well the employee works well in teams to ensure excellent results and a welcoming and inclusive workplace, which includes actively supporting and embracing diversity in its broadest representation. …

Often does not treat others with respect; uses an inappropriate tone and volume when talking to others or speaks to others in a way that does not show a recognition of their contribution to the team; has had incidents which have led to coaching and/or HR involvement in improving in this area.	Usually treats others with respect (e.g., giving others the opportunity to express their ideas/opinions; uses an appropriate tone and volume when talking to others; speaks to others in a way that shows a recognition of their contribution to the team); has not had incidents which have led to coaching and/or HR involvement in improving in this area.	Consistently treats others with respect (uses an appropriate tone and volume when talking to others; speaks to others in a way that shows a recognition of their contribution to the team).
Does not foster an inclusive environment; primarily seeks and accepts input from those they have worked with before or come from a similar background (area of expertise, education, etc.); is not open to opinions on how to do things differently.	Neither promotes or discourages an inclusive environment by actively supporting and embracing diversity in its broadest representation; accepts input from those who give it to them.	Promotes an inclusive environment by actively supporting and embracing diversity in its broadest representation, e.g., giving others the opportunity to express their ideas/opinions; seeks input from a wide variety of voices.
①	③	⑤

(Bobrow, 2021, personal communication)

3. Rater Reliability of BARS

- Annual employee performance evaluations
- Reliability for annual PEs and for exams
- Major BARS mystery
- Assumptions
- Selection decision stability
- A danger of z-score equating of boards
- Research needs

Annual Performance Evaluations

- BARS developed for annual evaluations
  - Much of published research is on annual PE
- Full year of behavior to remember and rate
  - Heavy cognitive load

BARS Reliability for Annual PEs

- .5-.7
  (Jacobs, Kafry & Zedeck, 1980, page 621)
- Mostly below .6
  (Schwab, Heneman & DeCotiis, 1975, page 557)
- .47 to . 73, median reliabilities for 58 munis
  (Landy, Farr, Saal & Freytag, 1976)
- .74 for college professors
  (Bernardin, 1977)

Comparison Reliabilities

- Personality factors
- Cognitive ability
- Forced-choice
- Non-behavioral rating scale

Personality Factors

- Cooperativeness: .72
- Sensitivity: .82
- Humility: .71
- Composure: .81
- Positivity: .84
- Awareness: .69

(Boyce & Capman, 2017)

Cognitive ability

- Wonderlic; .82 to .94.
  (Wonderlic, Inc., 2002, page 21)
- GATB, Alternate-Form: .76 to .91
  (Mellon Jr., et al., 1996, page 51)
- DAT: split-half largely in the .90s
  (Gregory, 2011, page 228)
- CAPS: alternate form .71 to .89
  (Knapp, Knapp & Michael, 1977, page 1083)

Forced-Choice

- .90 for 20 triads (with a total of 60 items)
  (Lepkowski, 1963, page 87)

Non-Behavioral Rating Scale

- .87 for engineer evaluation, 4 point scale for:
  - Capacity for growth - Engineering breadth
  - Engineering judgment - Drive vs cooperation
  - Creativity/Ingenuity - Quantity/quality of work
  - Leadership

  (Lepkowski, 1963, page 87)

Non-Behavioral Rating Scale

- .90, .91, .87, .81, 78 using 5 point Likert scales for:
  - Contributing to the team's work
  - Interacting with teammates
  - Keeping the team on track
  - Expecting quality
  - Having relevant KSAs

  (Ohland, et al., 2012, page 616)

BARS Reliability for Exams

- .23, .41, and .36 on **consecutive years** for ratings of oral communication of applicants to be entry-level police officers
- The authors suggest that these low reliabilities may be a result of using different raters across test administrations. (Hausknecht, Trevor & Farr, 2002)

BARS Reliability for Exams

- .83, .76 – communication skills
- .89, .69 – decision making skills
- .75, .65 – leadership skills
- First number is interrater reliability for group with “Frame of Reference” training and the other without such training.
  - (Schleicher, Day, Mayes & Riggio, 2002)

BARS Reliability for Exams

- .77, .67, .67, .65 for:
  - Situation assessment
  - Plan formation
  - Plan execution
  - Team learning
    (Georganta & Brodbeck, 2020)
  - (Not actually an exam, but was an exercise)

Inter and Intra-Board Reliability

- Oral board graded with BARS
- 2 questions, each rated by 2 boards (videos)
- Within board Spearman-Brown reliability estimates calculated (4 question-board combinations)
- Across board correlation (for each exercise)

Inter and Intra-Board Reliability

- Post-discussion ratings reported here
- Within board Spearman-Brown reliabilities
  - Exercise 1: .97, .91, for the two boards
  - Exercise 2: .97, .94 , for the two boards

Inter and Intra-Board Reliability

- Predict cross board correlations
- Product of .97 and .91 = .88 (Ex. 1)
- Product of .97 and .94 = .91 (Ex. 2)

Inter and Intra-Board Reliability

- Actual across board reliabilities
  - Exercise 1: .44
  - Exercise 2: .66

Inter and Intra-Board Reliability

- Unexpected discrepancy
  - Exercise 1: .88 v .44
  - Exercise 2: .91 vs .66
- Perhaps each individual board reached a consensus on the correct answer, but the consensus reached by the two boards differed somewhat.

Level of Inter-rater Reliability

- BARS reliability not really high for exams
- Intra board reliability > inter board

Desirable Levels of Reliability

- “If important decisions are made with respect to specific test scores, a reliability of .90 is the bare minimum, and a reliability of .95 should be considered the desirable standard.“
  (Nunnally & Bernstein, 1994, page 265)

Major BARS Mystery

- Low inter-rater reliability for annual PE
  - A full year's of performance to recall and rate
  - Raters may not have equal chance to observe
  - Ratees & raters may have personal relationships
- Exams do not have these limits
  - Short amount of performance to rate
  - Performance is recent
  - All raters have same data and are objective

Assumptions

- Reduce vagueness to reduce bias
- SMEs have clear, agreed-upon concepts
  - We can capture these by working with SMEs

4. Selection Decision Stability

- Practical vs statistical reliability
- Candidates are practical
- Psychometricans focus on reliability
- Candidates focus on who is selected
- Retest reliability vs selection reliability

Selection Decision Stability

- Large N exams
- Smaller N exams

Selection Decision Stability

- 3 variables drive stability in selections
  - Would same people be selected if retested
- Reliability
- Number of candidates
- Number of selections

Selections with Large N

- Assumptions
- 10,000 candidates
- 10% selected
- .9 or .7 reliability

Selections with Large N

- .9 reliability: 70% overlap in selections
  - On retesting 700 of the 1,000 would be selected
- .7 reliability: 47% overlap in selections
  - On retesting 470 of the 1,000 would be selected
- .6 reliability: 38% overlap in selections
  - On retesting 380 of the 1,000 would be selected

Selections with Modest N

- Assumptions
- 200 candidates
- 10% selected
- .9 or .7 reliability

Selection with Modest N

- On average, the same % agreement as large N
- More variability with smaller N
- Here are a few simulations
- .9 reliability, the overlap in selections:
  76%, 71%, 71%, 67%, 76%
- .7 reliability, the overlap in selections:
  35%, 57%, 67%, 43%, 38%

Many Board-Specific Selections

- Conclusion
- With typical levels of reliability, # of candidates, and # of openings
- 50% or more of the selections will depend on which raters comprise the rating board

5. Caution on Statistical Equating

- How to equate
- Assume raters grading on same criteria but some harder or easier graders

A Danger of z-Score Equating

- z-score equating creates distributions with equal means and standard deviations
- Dangers:
  - A superstar will end up looking ordinary
  - May not select the superstar
  - Candidates grouped with superstar will fare worse
  - Validity will go down

Research Needs

- How best to rate candidates who exhibit behaviors at several points on a BARS?
- Is watching candidate video multiple times helpful?
- What question characteristics are related to rater reliability?
- How to identify questions with answers that can be clearly described.

Research Needs

- How best to measure agreement in level
  - As opposed to simple correlation
- Are ratings after discussion more valid than before discussion?
- Does inclusion of a civilian rater increase the validity of a police promotional exam?

Research Needs

- Why is interrater reliability within a rater team higher than interrater reliability across rater teams (when video recordings are rated)?
- How to motivate raters to read the same sources as the candidates?

Research Needs

- Methods for objective scoring of complex answers (similar to essay answers)
- What level of correlation between dimensions is problematic?
- Can MC tests measure complex answers?
- How to measure halo?
  (Halo inflates reliability; Haladyna & Rodriguez, 2013)

6. Rater Training

- Professional Standards on Rater Training
- Other thoughts on training

Major Guiding Documents

- Joint *Standards* (formerly the APA *Standards)* (AERA, APA, NCME, 2014)
- SIOP *Principles for the Validation and Use* (SIOP, 2018)
- *Guidelines and Ethical Considerations for Assessment Center Operations*. (International Taskforce on Assessment Center Guidelines, 2015).

Standards on Fairness

- "fairness is a fundamental issue for valid test score interpretation, and it should therefore be **the goal** for all testing applications."

  (AERA, APA, NCME, 2014, page 62, col 2, par 3)
- Are board-specific selections unfair?

Standard 6.9

Standard
Those responsible for test scoring should establish and document quality control processes and criteria. **Adequate training should be provided. The quality of scoring should be monitored and documented**. Any systematic source of scoring errors should be documented and corrected.

Standard 6.9

Comment

Criteria should be established for acceptable scoring quality. Procedures should be **instituted to calibrate scorers** (human or machine) prior to operational scoring, and to **monitor how consistently scorers** are scoring in accordance with those established standards during operational scoring.

Standard 4.20

Standard

The process for … training, qualifying, and monitoring scorers should be specified… should result in a degree of accuracy and agreement … describe processes for assessing scorer consistency and potential drift over time in raters' scoring.

Standard 4.20

Comment

... The basis for determining scoring consistency (e.g., percentage of exact agreement, percentage within one score point, or some other **index of agreement**) should be indicated. Information on scoring consistency is essential to estimating the precision of resulting scores.

Principles

- "If raters are an integral part of the selection procedure, as in some work samples, then the **reliability and agreement** of their ratings should be determined and documented."
  (SIOP, 2018, page 34)
  (emphasis added)

Guideline 7

- "Assessor Training—Assessors must receive thorough training and **demonstrate** performance that **meets pre-specified criteria.**"
  (emphasis added)

Other Ideas on Rater Training

- If want to change focus of department, select raters from departments that you emulate
- Demanding to observe and recall, so have raters watch videos twice
  (Schleicher et al 2002, page 736, col 2, par 2)

Other Thoughts on Training

- Help raters reach agreement on correct answer
- Refine answer key after viewing candidate responses.

7. Why Are BARS Widely Used?

- BARS expensive (i.e., hard work)
- BARS not clearly better than other scales (Cascio & Aguinis, 2011, page 95)
- Forced choice abandoned by military because raters could not tell what rating they were giving, so they tried to beat the system (Dickenson & Zellinger, 1980, page 153)

Why Are BARS Widely Used?

- Face validity
- Easy to explain
- Candidates like concrete feedback of BARS
- More amenable to rater training
- Can focus on job tasks rather than traits
- Appear fair

(Debnath, Lee & Tandon, 2015)

Q&As at End of Session

- Please type questions in chat
  - We will try to address chat questions after the last presenter
- Feel free to contact me at any time about this topic
  - (617) 244-8859
  - jpw@jpwphd.com

References

- AERA, APA, NCME. (2014) *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

References

- Bernardin, J. H. (1977). Behavioral Expectation Scales Versus Summated Scales: A Fairer Comparison. *Journal of Applied Psychology*, *62*, 422-427.

References

- Boyce, A. S. & Capman (2017) ADEPT-15® *Technical Documentation; Development and Validation of Aon Hewitt's Personality Model and Adaptive Employee Personality Test (ADEPT-15).* Author.

References

- Casio, W. F. & Aguinis, H. (2011). *Applied Psychology in Human Resource Management*. Boston: Pearson.

References

- Debnath, S. C., Lee, B. B. & Tandon, S. (2015). Fifty Years and Going Strong: What Makes Behaviorally Anchored Rating Scales So Perennial as an Appraisal Method*? International Journal of Business and Social Science*, *6*, 16-25.

References

- Dickinson, T. L. & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied Psychology*, *65*, 147-154.

References

- Georganta, E. & Brodbeck, F. C. (2020) capturing the four-phase team adaptation process with behaviorally anchored rating scales (BARS). *European Journal of Psychological Assessment*, *36*, 336–347.

References

- Gregory, R.J. (2011).*Psychological testing history, principles, and applications* (6th ed.) Boston, MA: Allyn & Bacon.

References

- Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating Test items.* New York: Routledge.

References

- Hausknecht, J. R., Trevor, C. O. & Farr, J. L. (2002). Retaking ability tests in a selection setting: implications for practice. Effects, training performance, and turnover. *Journal of Applied Psychology*, *87*, 243–254.

References

- International Taskforce on Assessment Center Guidelines. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *41*, 1244–1273.

References

- Jacobs, R., Kafry, D. & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psvchology*, *33*, 595-640.

References

- Kell, H. J., Martin-Raugh, M. P., Carney, L. M., Inglese, P. A., Chen, L. & Feng, G. (2017*). Exploring Methods for Developing Behaviorally Anchored Rating Scales for Evaluating Structured Interview Performance; Research Report ETS RR–17-28*. Princeton, NJ: Educational Testing Service.

References

- Knapp, R. R., Knapp, L. & Michael, W. B. (1977). Stability and concurrent validity of career ability placement survey (CAPS) against the dat and the GATB. *Educational and Psychological Measurement*, *37*, 1081-1085.

References

- Landy, F. L. (1977). *Performance appraisal in police departments*. Arlington, VA: Police Foundation.
- Landy, F. J., Farr, J. L., Saal, F. E. & Freytag, W. R. (1976). Behaviorally Anchored scales for rating the performance of police officers. *Journal of Applied Psychology*, *61*, 750-758.

References

- Maurer, S. D. (2002). A Practitioner-based Analysis of Interviewer Job Expertise and Scale Format as Contextual Factors in Situational Interviews. *Personnel Psychology*, *55*, 307-327.

References

- Kell, H.J., Martin-Raugh, M.P., Carney, L.M., Inglese, P.A., Chen, L., & Feng, G. (2017). Exploring Methods for Developing Behaviorally Anchored Rating Scales for Evaluating Structured Interview Performance. ETS Research Report No. RR-17-28. Princeton, NJ: Educational Testing Service.

References

- Lepkowski, J. R. (1963). Development of a forced-choice rating scale for engineer evaluation. *Journal of Applied Psychology*, *47*, 87-88.

References

- Mellon Jr., S. J., Daggett, M., MacManus, V. & Moritsch, B. (1996) *Development of General Aptitude Test Battery (GATB)* Forms E and F. Sacramento, CA: Pacific Assessment Research and Development Center.

References

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

References

- Ohland, M. W., Loughry, M. L., Woehr, D., Bullard, L. G., Finelli, C. J., Layton, R. A., Pomeranz, H. R. & Schmucker, D. G. (2012). The Comprehensive Assessment of Team Member Development of a Behaviorally Anchored Rating Scale for Self- and Peer Evaluation. *Academy of Management Learning & Education*, *11*, 609-630.

References

- Pulakos, E. D. (1997). Ratings of Job Performance. In D. L. Whetzel & G. R. Wheaton (Eds.) *Applied Measurement Methods in Industrial Psychology*. Palo Alto, CA: Davies-Black.

References

- Schleicher, D. J., Day, D. V., Mayes, B. T. & Riggio, R. E. (2002). *Journal of Applied Psychology*, *87*, 735–746.
- Schwab, D. P., Heneman, H., III & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: a review of the literature. *Personnel Psychology*, *28*, 549-562.

References

- SIOP (2018*). Principles for the Validation and Use of Personnel Selection Procedures, 5th ed*. Bowling Green, OH: Author.

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

- Wonderlic, Inc. (2002). *Wonderlic Personnel Test & Scholastic Level Exam User's Manual*. Libertyville, IL: Author.