Assessing Performance in Clinical and Other Rater-Mediated Scenarios

Kenneth Royal, PhD, MSEd

Session Outcomes

At the end of this session, you will be able to:

- Recognize key concepts in the measurement of performance
- Recognize challenges that complicate measurements of performance
- Identify sources of error/biases in performance assessments
- Conceptualize the key concepts underpinning standard setting
- Evaluate complexities in defining levels of performance
- Evaluate complexities in determining thresholds/cutpoints for various levels of performance
- Appreciate the science behind good performance assessment

Ken’s Goal for this Session

- To crowdsource (concrete) ideas about how we can improve performance assessments in the CVM
- Identify changes we need to make to create more objective measurements of performance
- Leave this room with some “first step” ideas in place

Let’s Start with an Exercise in Measurement

Questions from Exercise

- How consistent was your set of measurements? (*intra-rater reliability*)
- How consistent were the measurements between members in your group? (*inter-rater reliability*)
- Why do you think there were inconsistencies within your set of individual measurements?
- Why do you think there were inconsistencies in measurements between members of your group?

A Daunting Challenge

- Conducting error-free measurement is challenging even with physical measurements (measuring height, blood pressure, etc.)
- Conducting error-free measurement in educational measurement is impossible
- However, imperfect measurement doesn't mean we settle and make no effort to create "more objective" assessments
- Think about how much more complicated measurement becomes when we try to measure latent traits (e.g., ability, attitude, knowledge)
- Learners should not be advantaged or disadvantaged based on who is evaluating them
- We must have processes in place to ensure measurements are as accurate, and error-free, as possible

A Daunting Challenge

- When we evaluate learners' performance, we are making measurements
- The challenge:
  - Applying your "ruler" fairly and consistently
  - Developing mindshare with other evaluators
  - Applying your collective rulers fairly and consistently
  - Doing so with minimal error/bias

Questions

1) How can you ensure you apply your “ruler” fairly and consistently?

2) How can you develop mindshare (a common lens) with other evaluators?

3) How can you and your colleagues collectively apply your rulers fairly and consistently?

4) How can you do this with minimal error/bias?

Rater Errors and Biases

Exercise

Exercise consists of two rounds:

Round 1:

1) Consider each type of rater error and place an “X” beside any error you have previously committed in the “Rd 1” column
2) Count how many of the errors have you committed and note them on the bottom of page 2
3) Are there any errors that you will likely no longer commit now that you are consciously aware of the type of error (e.g., conscious bias)?

Exercise

Round two:

1) Revisit each of the errors you previously flagged and ask yourself “how many of these errors will likely be a persistent challenge for me?”
2) Place an “X” beside errors that you anticipate will remain a challenge
3) Count how many errors will likely remain a persistent challenge and note them on the bottom of page 2
4) Take a few moments to consider what you, colleagues and/or the CVM in general can do to mitigate the errors that awareness alone is unlikely to resolve
5) Note those ideas in the open text box at the bottom of page 2

How to Address these Challenges

So, today we have already…

- Made the connection between physical measurement and social/behavior measurement (“know our role”)
- Identified/confronted some of the many biases we may encounter when rating performance
- Elsewhere, we/others have developed, and are continuing to develop standards (e.g., EPAs)
- Elsewhere, we/others have developed, and are continuing to develop instruments (e.g., rubrics, checklists, etc.)

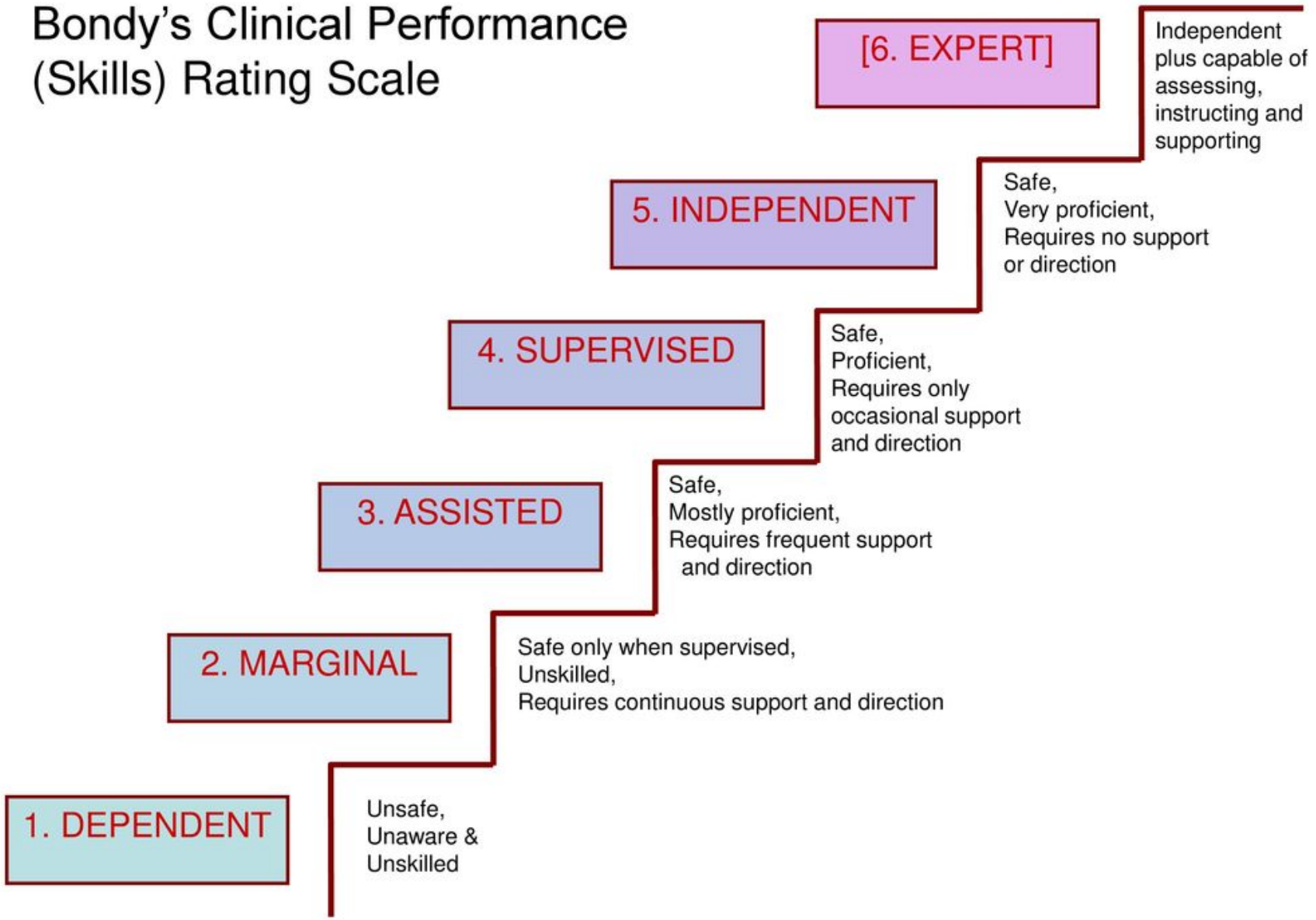
Now, let’s talk a bit about “standard setting” (what we need to do once standards are articulated)

Standard Setting

Critical Considerations

- Simply put, “standard setting” is the process of establishing one or more cut scores on an assessment (e.g., Pass/Fail standard on the NAVLE), or other continuum of events
  - Speed limit
  - Height required to ride amusement park rides
- In performance assessments, we sometimes have pass/fail situations
- Usually, however, there are multiple levels of performance
- In short, standard setting is used as a scientifically acceptable approach to determine where the thresholds that differentiate levels of performance lie

Bondy's Clinical Performance (Skills) Rating Scale



Critical Considerations

- Standard setting demands that we use scientifically acceptable procedures that lead to a decision or result that is fundamentally fair and reasonable
- Obviously, just as equally qualified and interested persons could disagree about whether a procedure is systematic and rational, so too might reasonable persons disagree about whether the results of any particular standard-setting process are fundamentally fair
- The notion of fairness is, to some extent, subjective and necessarily calls into play persons' preferences, perspectives, biases, and values
- We must accept that subjectivity exists, yet utilize a systematic set of rules/procedures, informed by measurement science, to create cut points/thresholds that are fair and reasonable

Standard Setting

10 Generic Steps in Setting Performance Standards

#1 - Select a large and diverse panel

#2 - Choose an appropriate standard-setting method; prepare training materials and standard setting meeting agenda

#3 - Prepare descriptions of the performance categories

#4 - Train participants to use the standard setting method

#5 - Compile item ratings or other judgments from participants and produce descriptive/summary information or other feedback for participants

#6 - Facilitate discussion among participants of initial descriptive/summary information

10 Generic Steps (cont.)

#7 - Provide an opportunity for participants to generate another round of ratings; compile information and facilitate discussion as in Steps 5 and 6

#8 - Provide for a final opportunity for participants to review information; arrive at final recommendation performance standards

#9 - Conduct an evaluation of the standard setting process, including gathering participants' confidence in the process and resulting, performance standard(s)

#10 - Assemble documentation of the standard setting process and other evidence, as appropriate, bearing on the validity of resulting performance standards

Two Examples of Standard Setting

Angoff Process

Involves systematically combining subjective judgments of content experts

***Steps*:**
1. Discuss and internalize the ability of a borderline examinee
2. Review test questions and evaluate the difficulty of each
3. Estimate the proportion of borderline examinees who will answer each question correctly

Angoff Method

	Judge A	Judge B	Judge C	Judge D	Item Avg.
Item 1	0.59	0.52	0.46	0.60	0.54
Item 2	0.63	0.72	0.78	0.63	0.69
Item 3	0.54	0.45	0.41	0.58	0.50
Item 4	0.68	0.65	0.45	0.57	0.59
Item 5	0.97	0.80	0.98	0.74	0.87
Item 6	0.68	0.60	0.52	0.54	0.59
Item 7	0.83	0.83	0.95	0.75	0.84
Item 8	0.64	0.60	0.80	0.70	0.69
Item 9	0.68	0.56	0.72	0.51	0.62
Item 10	0.58	0.73	0.64	0.51	0.62
Judge Avg.	0.68	0.65	0.67	0.61	**0.653**

Bookmarking Process

- Involves the use of an ordered item booklet (items are ordered from least difficult to most difficult)
- Expert panelists are instructed to determine if a minimally competent candidate (MCC) has at least a two-thirds probability of answering an item correctly
- Panelists review the increasingly difficulty items, then place a bookmark where they believe this threshold (the point where a MCC no longer has a 67% probability of answering correctly) occurs
- Panelists scores are calculated

Issue	Consequence	Measure*	SE**
Taking a graded quiz or examination for another student	Suspension	99.89	5.10
Changing a response after a paper/exam/quiz was graded, then reporting that there had been a misgrade and requesting credit for your altered response	Failure of Course	88.72	4.08
Using unauthorized cheat sheets or other materials during a quiz or examination		82.00	3.59
Copying from another student during a quiz or examination		79.93	3.46
Claiming to have handed in a paper/examination when in reality you did not		70.98	2.96
Permitting another student to look at your answer sheet during a quiz or examination.		70.62	2.95
Removing items from a reserved reading file so that others will not have an opportunity to review them	Failure of Assignment	59.66	2.49
Asking another student for the questions and/or answers to an examination which he/she had taken and you will take in the future		55.72	2.36
Providing information about an exam that was intended to be confidential.		53.67	2.30
"Making up" sources for bibliographic citation		46.51	2.13
Posting unauthorized information about exams, assignments, quizzes, etc. on social media		45.21	2.10
Using a false excuse to postpone an exam		44.33	2.07
Using direct quotes from other sources without giving proper reference		44.15	2.07
Listing false completions on your online clinical skills completion summary		38.23	1.96
Working with another student on a quiz or homework assignment that was assigned as individual work.	Letter Grade Decrease on Assignment	32.75	1.88
Presenting your clinical skills book for signing without actually completing the skill		32.70	1.87
Using unauthorized test questions from a previous year, including materials found on public websites.		29.81	1.84
Claiming to have attended class when you actually did not	Written Warning	27.70	1.82
Visiting a professor after an exam or at the end of the semester to bias his/her grading		22.17	1.77
Listing unread sources in the bibliography of an assignment		19.80	1.76
Missing class or lab due to a false excuse		17.47	1.75
Doing less than your fair share in a group project or a laboratory	Verbal Warning	8.44	1.76
Failing to prepare adequately for a group assignment or laboratory		-0.05	1.80

Where Do We Go From Here?

- How can we ensure all raters in the CVM conduct performance assessments in a fair and consistent manner?
  - Online training modules?
  - “Certified clinical evaluators”?
  - Simplify our instruments (e.g., rubrics)?
  - Simply our grading schema (e.g., pass/fail; meets standard / does not meet standard)?
- What will it take to get your (service) team on board?
- What challenges might you anticipate when attempting to develop mindshare with your team?

Where Do We Go From Here?

- What will you do when experts disagree? How will you reconcile differences in expert opinions?
- What might a well-trained service team look like?
- How frequently would raters need to re-calibrate themselves? Their teams?
- What can we do to make the faithful execution of calibrated ratings a cultural norm among your team/service?

Questions?

Please contact me (kdroyal2@ncsu.edu) if you would like to discuss all things assessment, evaluation, measurement and education research