The Rater Bias Inventory: A Tool for Discerning Conscious and Unconscious Bias

Kenneth Royal, PhD, MSEd
North Carolina State University

Overview

Minimizing the influence of rater errors is a considerable challenge in performance assessment contexts. Royal (2018) presented a list of 45 common errors that impede quality measurement in medical and health professions education. This adaptation of the original work is intended to offer raters a tool (the Rater Bias Inventory, RBI) to identify potential biases that may influence an individual's score by introducing unwanted measurement error.

Directions

Self-Assessment

Individuals wishing to identify biases that may potentially result in rater errors are encouraged to perform the following steps:

1) Begin by reviewing each rater error/behavior and place an "X" beside any error you have previously committed in the "Round 1" column.
2) Count how many errors you have previously committed and note the number in the "Total Count" box at the bottom of page 3.
3) Take a moment to reflect on your responses. Are there any errors that you will likely no longer commit now that you are consciously aware of the error (e.g., conscious bias)?
4) For Round 2, review the errors you previously flagged in Round 1 and identify any errors that likely will persist despite being consciously aware. Place an "X" beside any error that likely will remain a challenge for you.
5) Count how many errors you identified in Round 2 and note the number in the "Total Count" box at the bottom of page 3.
6) Take a few moments to consider what you can do to mitigate the errors that awareness alone is unlikely to resolve and note your ideas in the open text box at the bottom of page 3.

Collaborative (Team-Based) Assessment

Individuals working with teams (e.g., clinical department, standardized patients, etc.) might find the RBI useful as part of a rater training ("calibration") exercise. Teams using the RBI for this purpose might have each member complete the RBI independently, then compare results. Dialogue might involve identifying any errors that are common to multiple individuals and strategies that may be employed to mitigate various errors.

Reference

Royal KD. Forty-five common rater errors in medical and health professions education. Educ Health Prof, 2018(2)1:33-5. Available from: http://www.ehpjournal.com/text.asp?2018/1/2/33/251905.

Round 1	Round 2	Types of errors	Rater behavior
		Assimilation effect	Intentionally providing ratings that likely will be similar to other raters in an effort to avoid appearing 'extreme'
		Carryover effect	Allowing the performance of a prior individual to affect (either positively or negatively) one's rating
		Central tendency	Avoiding extreme scores and rating most individuals as average
		Changing times	Providing ratings that are influenced by generational differences
		Cheerleader effect	Providing higher scores, regardless of performance, to support those individuals being assessed or evaluated
		Clashing standards	Providing lower scores because the standards addressed by the instrument differ from the personal standards of the rater
		Clashing values	Providing lower scores because the values expressed are inconsistent with those of the rater
		Contrast effect	Providing ratings that intend to compare individuals rather than rate performance relative to a standard
		Cowardice	Providing a higher score due to fear of being challenged or retaliated against
		Different-from-me effect	Assigning lower scores to individuals who possess different qualities or attributes to that of the rater
		Distraction	Providing a rating for an observation that occurred when the rater was distracted for some reason
		Drift	Providing ratings that become increasingly inconsistent with one's previous ratings
		Explicit bias	Providing ratings (both favorable and unfavorable) based on a conscious bias
		Extremism	The tendency to exclusively assign ratings at the extreme ends of the rating scale (middle category avoidance)
		Fatigue	Providing a questionable rating as a result of feeling tired
		Favoritism	Providing higher scores because an individual is well-liked by the rater and/or his/her colleagues
		First impression	Providing a judgment that is based solely on an initial impression
		Frustration	Allowing one's feelings of frustration/anger to affect one's ratings
		Implicit bias	Providing ratings (both favorable and unfavorable) based on an subconscious bias
		Impressiveness	The tendency for a rater to experience feelings of admiration or awe that clouds his/her judgment
		Halo effect	Providing a higher score as a result of giving an individual the benefit of a doubt
		Horn effect	Providing a lower score as a result of allowing one trait to overshadow others
		Hurriedness	Providing ratings that are influenced by one's desire to quickly complete the task
		Gestalt phenomenon	Providing a score that is based on an overall impression (rater is unable to differentiate various aspects of performance)
		Guilt by association	Providing a lower score to an individual because of his/her association with someone else
		Insecurity	Providing a higher score due to fear of being unable to defend a lower score
		Length	Consistently providing a better score for longer or shorter performances
		Leniency	Providing ratings that more favorably represent an individual's actual performance

Cont.

Round 1	Round 2	Types of errors	Rater behavior
		Personality clash	Lowering a score because the rater has a fundamental incompatibility in personality with the individual being assess
		Primacy	Focusing only on the earliest observation(s) and ignoring more recent observations
		Rater competence	Failing to select "N/A", "not observed", "unable to judge", etc. when such a response would be appropriate.
		Recency	Focusing only on the most recent observation and ignoring all previous observations
		Repetition factor	Altering a score due to having seen the same or similar performance multiple times
		Retaliation	Providing a lower score as an act of revenge
		Scale interpretation	Mistakenly perceiving a rating to mean something different than it intends
		Score range restriction	Failing to differentiate individuals on the behaviors (or latent trait) being measured.
		Self-scoring	Making a judgment that reads too much into a performance
		Similar-to-me effect	Awarding better ratings to individuals who share similar qualities or attributes to that of the rater
		Single deficiency focus	Providing a lower score as a result of overly focusing on a single deficiency
		Skimming	Failing to fully consider an entire performance by instead focusing only on limited portions of a performance
		Sudden death	Providing a lower score because some aspect of a performance invokes a negative rater response
		Sympathy score	Providing a higher score based on sympathy for the individual being assessed
		Severity	Providing ratings that are unduly harsh or critical
		Timeline	The tendency to provide ratings that exceed the bounds of a specified timeline
		Trait	Attributing too weight to one of several important aspects (e.g., communication, appearance, etc.) of performance
Total Count	**Total Count**		

What can/will you do to mitigate the errors you identified above? (Enter your notes here)

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____