ASSOCIATION OF STANDARDIZED PATIENT EDUCATORS

Transforming professional performance through the power of interaction

Subjective Variables in Video-Based Quality Assurance of Communication Ratings

By Tony Errichetti, National Board of Osteopathic Medical Examiners

When I recently left my position at a medical school to join the National Board of Osteopathic Medical Examiners, I thought my debriefing days were over. Instead I find debriefing is the main tool I use whenever I ask SPs to explain their communication ratings as part of our on-going training and quality assurance (QA) process. Here are some thoughts about the value and limitations of video-based QA of communications ratings and the role of debriefing in maintaining accuracy.

Rating communication on a global / holistic scale is challenging for SPs and performance raters. Rater bias, carelessness, and misinterpretation of the scoring rubric are potential sources of rating error.[1] QA methods, such as double scoring by video, help ensure the fairness and accuracy of SP scores. However, external raters of SP accuracy – trainers, SPs, etc. – are subject to the same scoring errors. Additionally, qualitative differences between the video and in-room experience, and the judgments involved in assessing communication, can generate scoring disagreements. A debriefing conversation can determine how the respective ratings were developed to begin with: *"I scored X, but you scored Z. Let's compare notes."*

Rating accuracy is determined during such good faith conversations in which both SP and QA rater give evidence for their scoring and determine who is "more" accurate. The QA rater, even if he or she is the one who trained the SP, may be in error: scoring discrepancies are often legitimate disagreements and should not be broached as automatic judgments of the SP's accuracy. And scoring disagreements may be a by-product of the QA process itself.

The following is a short list of subjective variables that can account for scoring decisions and disagreements when humans are involved in the rating process.

The Rashoman Effect, or multiple interpretations of the same event by different people, is a reflection of our human nature.[2,3] Checklists are concrete (yes/no,

happened/didn't happen) and relatively easier to score and verify. But communication rating scales require nuanced judgments. Question-and-Answer (QA) debriefing examines the nature of those judgments and the degree to which ratings are based on behavioral definitions and anchors.

"Empathy is…"
"Some of the expected empathic behaviors include…"
"I rated the examinee this way because…"

Well-trained raters, whether SPs or trainers, will not always agree on what they observe. I accept the Rashoman Effect as a starting point of a QA debriefing review. Indeed, we cannot standardize our impressions and observations beyond a certain level of tolerance. In an SP exam, a sufficient number of stations can level out rater differences. The multi-station exam form, in part, reflects our acceptance of the limitations of human perception, while providing a variety of patient presentations for the student to show what he or she is minimally capable of in a clinical setting.

The SP and/or QA rater can have different understandings of the scoring system. This is why discussing scoring differences is important. The one who is most accurate is the one who has the best explanation for their scores within the context of the case. Do both the QA rater and the SP have the same working definition of the communication dimensions to be scored? Are the behavioral anchors used to guide scoring clear and explicit? Even if the SP and QA rater agree in their scoring the rationale can be flawed, with agreement occurring by chance. Therefore it's important to review all scores, and not simply those where there is a disagreement. Debriefing communication scoring is like a forensics exercise in which both raters give evidence for their scoring decisions.

Live/in-room ratings and QA video ratings are "apples and oranges." When scoring live in the room, the SP performs, responds and interacts with an examinee. Post-encounter they score holistically from memory but are arguably in the best position to assess communication subtleties of "softer" skills, e.g. non-verbal responses. When

raters score from video they use a different cognitive process. They can stop-start-reverse-replay the video. It becomes an analytic vs. a holistic exercise, the

microanalysis of an interaction. Sometimes SPs, when re-scoring their own videos at a later date, disagree with their in-room ratings. When debriefed about the differences they often say "it just looked different on video." It is indeed a different perceptual experience. For a fascinating book comparing holistic vs. analytic decision making, read Malcolm Gladwell's *Blink: The Power of Thinking Without Thinking* (Boston: Back Bay Books, Little, Brown, 2005, ISBN 0-316-17232-4).

Video as Tool and Lens. Video imposes physical conditions and limitations on QA. When done on a PC the video image is relatively small. The sound quality even with headphones may not be optimal. When done for extended periods of time viewers can experience "computer vision syndrome."[4] This is caused by screen glare, image size, viewer distance from the screen and posture. And it's a tedious, exacting process. SPs who like and prefer QA work I suspect have an aptitude for air traffic control.[5] But because video rating provides a remove, the QA rater can focus on the more concrete communication behavioral anchors such as speech and gross body language. Such analytic observations are extremely important when comparing scores. Debriefing can add meaning to the observations.

I welcome the opportunity to discuss these issues with the SP educator community.

Bio: Tony Errichetti, PhD, is Director of Doctor-Patient Communication Assessment at the National Board of Osteopathic Medical Examiners, National Center for Clinical Skills Testing. tony.errichetti@gmail.com or aerrich@nbome.org

---

1. Boulet JR, McKinley DW, Whelan GP, Hambleton RK (2003). Quality assurance methods for performance-based examinations. *Adv Health Sci Educ Theory Pract,* 8(1):27-47.

2. Davenport, Christian (2010). "Rashomon effect, observation, and data generation". Media Bias, Perspective, and State Repression. Cambridge University Press. p. 55.

3. Karl G (Match 1988). The Rashoman effect: when ethnographers disagree. *American Anthropologist* 90 (1): 73.81.

4. American Optometrics Association website: http://www.aoa.org/patients-and-public/caring-for-your-vision/protecting-your-vision/computer-vision-syndrome?sso=y

5. 13 characteristics of an air traffic controller. http://waynefarleyaviation.com/2010/09/13-characteristics-of-an-air-traffic-controller/