On Rating Angels: The Halo Effect and Straight Line Scoring

Jonathan Sherbino, MD, MEd
Geoff Norman, PhD

The halo effect is "a tendency to grade for general merit at the same time as for the qualities, and to allow an individual's general position to influence his position in the qualities."[1] It has a long and storied history, particularly for anyone supervising students in clinical settings or, for that matter, for anyone who must rely on subjective ratings of individuals for promotion decisions. We suspect most supervisors and program directors can recall multiple examples of the extreme in this regard, such as the long, consistent line of 8 out of 10 checks without deviation on the rating form. This phenomenon is called *straight line scoring* (SLS) in the study by Beeson et al[2] in this issue of the *Journal of Graduate Medical Education*.

The halo effect is ubiquitous. As early as 1920, Thorndike[3] identified it in teacher ratings, causing him to bemoan the fact that raters are:

> "Unable to treat an individual as a compound of separate qualities and to assign a magnitude to each of these in independence of the others.. . . The halo . . . seems surprisingly large, though we lack objective criteria by which to determine its exact size."[3(pp28,29; quoted in 1)]

Nothing in the subsequent century has happened to change this phenomenon. Although some researchers have hypothesized that the phenomenon is a consequence of insufficient specification of the individual dimensions, the *halo* appears to persist, despite multiple permutations to the form (ie, type and number of questions) and scale. Indeed, some researchers appear blissfully unaware of the effect. When an article reports that the internal consistency was high (> 0.9), it does not appear to dawn on the investigator that this is simply too high in a situation where the rater is supposed to be rating separable and independent categories.[4]

For example, Tromp et al[5] had raters assess 19 competencies over successive 3-month intervals. They found that "scores . . . show[ed] excellent internal consistency ranging from .89 to .94."[5(p199)] This may not be excellent at all, but rather evidence that individual competencies cannot be distinguished. In another example, Ginsburg et al[6] studied the ratings of 63 residents over 9 rotations with a 19-item instrument, where a factor analysis showed a single *g* factor accounting for 66% of the variance.

In a study of more than 1800 evaluations of 155 medical students, differences among students accounted for 10% of the variance, and differences among raters accounted for 67%.[7] A factor analysis produced a single factor explaining 87% of the variance, a rating more telling of the supervisor than the student.[7]

Even in the rating of psychomotor skills using direct observation, the halo effect appears to be endemic. A systematic review by Jelovsek et al[8] showed internal consistency alpha coefficients of 0.96,[9] 0.89,[10] 0.91–0.93,[11] 0.97,[12] 0.97,[13] 0.98,[14] and 0.97.[15] It seems implausible that items like "use of radiographs," "time and motion," and "respect for tissue" would be so highly correlated on a single 10-item instrument to yield an alpha of 0.97.[15] All of this suggests that SLS is merely the tip of the iceberg in defining the limits of observer judgment when scoring multiple items consecutively.

The halo effect has at least 2 components and potentially multiple causes. First, ratings generally are too high and skewed to the positive end of the scale. Second, the general score or several key subdomains that inform the general score dictate a consistently high score in all other subdomains. The term *halo* is apt; not only are all the ratings illuminated by an overall light, but the light is in the form of a warm glow. The *millstone effect* is the opposite phenomenon—a low general rating that weighs down a consistent low score of all other subdomains. However, this effect is uncommon in both education practice and literature.

Why does the halo effect arise? Cooper[1] reviewed a number of potential causes, although he does not estimate the relative contribution of each. *Undersampling* reflects an inadequate number of observations, which may well apply in the context of first-time ratings as suggested by Beeson et al.[2] *Engulfing* occurs when ratings of individual attributes are

DOI: http://dx.doi.org/10.4300/JGME-D-17-00644.1

influenced by an overall impression. Beeson et al[2] propose that the final ratings of graduating residents may have been a result of this phenomenon, where the score was linked to predicted success of graduation. *Insufficient concreteness* results from an inadequately precise definition or demarcation of subdomains. Perhaps we should anticipate a halo effect base rate in assessing milestones, as many milestones are informed by competencies that are common across them. What is not included in Cooper's review is rater attention, where long assessment forms promote rater fatigue and pragmatic SLS solutions.

We propose a simpler explanation—differentiation and integration. Differentiation occurs when an observation is turned into a judgment. The factor analyses of summary judgments reviewed suggest that people are capable of differentiating 1 domain and, occasionally, 2 domains. Residency programs that require a supervisor to rate a resident on multiple domains at the end of a day (or even worse, at the end of a rotation when elapsed time weakens recall) will impair the supervisor's ability to differentiate between multiple domains. Integration comes into play at the point of creating a summary judgment, assembled from the aggregated judgments based on the observations of performance. The consequence of these 2 phenomena is what is observed as the halo. All ratings of ostensibly independent (or nearly so) attributes are highly correlated, and with rare exceptions center on an above average value.

Straight line scoring represents the pathological extreme of this phenomenon, amounting to a perfect correlation among domains. Beeson et al[2] found that this arose in only about 6% of all assessments. However, it was not evenly distributed across programs, with 4% of programs submitting more than 50% of their assessments with SLS. As a quality assurance process, such programs should undergo evaluation to better understand how the assessment processes can be improved.

Unique to the work of Beeson et al[2] is the description of halo effect with group ratings done by Clinical Competency Committees (CCCs). In the current design, the CCC is somehow expected to add up all observations from the past 6 months, using a complex mental calculus that discounts early observations and weighs later ones more, while taking into account the context in which they occurred, and so on. The psychology of group rater cognition and groupthink is not even included in this challenge. Why is this worrisome? CCCs play a key role in competency-based graduate medical education. The Accreditation Council for Graduate Medical Education (ACGME) requirements expect CCCs to (1) integrate assessment data into milestone scores, and (2) provide a summary judgment about resident suitability for unsupervised practice (ie, differentiate).[16] Yet, the work of Beeson et al[2] suggests these 2 tasks may be confounded by halo.

What to do? First, we suggest the adoption of programmatic assessment across residency programs: multiple observers, sampling *single* sentinel competencies within a systematic integrated whole.[17] A systematic approach to ensure multiple biopsies of performance of a single task may improve differentiation and lessen the influence of 1 milestone on others.

Second, we suggest the adoption of an actuarial model (eg, regression analysis) to provide a summary judgment (ie, differentiation) of a resident's global performance.[18] The jury model of decision-making, used by CCCs, is prone to the halo effect. An actuarial model allows the weighting of variables by clinical content experts. The relations among variables can be calculated without subjugation to the psychological phenomena that give rise to the halo effect, while providing a final statistical measure of performance. For example, the CCC could weigh the importance of all milestones, and suggest inclusion of other assessment data, such as in-training examinations. A regression analysis of the combined data would provide a summary judgment about resident performance.

Third, we suggest reconstituting CCCs to focus on oversight of resident assessment; provision of tailored educational prescriptions to residents, based on programmatic assessment data; and maintenance of accountability for the summary judgments produced by the above processes.

As Beeson et al[2] showed, ACGME-accredited emergency medicine residency programs demonstrated an SLS rate of approximately 6%. Is the frequency of this severe form of halo effect worrisome? For the minority of programs with a rate of SLS greater than 50%, there is clearly cause for concern. However, SLS is likely the tip of the iceberg. The real question is the extent to which these assessment processes are compromised by less severe halo effects.

The assumption that competence amounts to achievement of multiple independently assessable abilities, which may be mastered at different rates by different residents, is central to competency-based medical education. Strategies based on summative ratings, whether by individual supervisors or committees, inevitably lead to the halo effect. We must devise alternative strategies that permit valid assessment at the level of individual competency.

References

1. Wells FL. A statistical study of literary merit with remarks on some new phases of the method. *Arch*

Psychol. 1907;1(1):5–30. Cited in: Cooper WH. Ubiquitous halo. *Psychol Bull*. 1981;90(2):218.
2. Beeson MS, Hamstra SJ, Barton MA, et al. Straight line scoring by clinical competency committees using emergency medicine milestones. *J Grad Med Educ*. 2017;9(6):716–720.
3. Thorndike EL. A constant error in psychological ratings. *J Applied Psychol*. 1920;4:25–29.
4. Norman G. Data dredging, salami-slicing, and other successful strategies to ensure rejection: twelve tips on how to not get your paper published. *Adv Health Sci Educ Theory Pract*. 2014;19(1):1–5.
5. Tromp F, Vernooij-Dassen M, Grol R, et al. Assessment of CanMEDS roles in postgraduate training: the validation of the Compass. *Patient Educ Couns*. 2012;89(1):199–204.
6. Ginsburg S, Eva K, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med*. 2013;88(10):1539–1544.
7. Sherbino J, Kulasegaram K, Worster A, et al. The reliability of encounter cards to assess the CanMEDS roles. *Adv Health Sci Educ Theory Pract*. 2013;18(5):987–996.
8. Jelovsek JE, Kow N, Diwadkar GB. Tools for the direct observation and assessment of psychomotor skills in medical trainees: a systematic review. *Med Educ*. 2013;47(7):650–673.
9. Laeeq K, Infusino S, Lin SY, et al. Video-based assessment of operative competency in endoscopic sinus surgery. *Am J Rhinol Allergy*. 2010;24(3):234–237.
10. Syme-Grant J, White PS, McAleer JP. Measuring competence in endoscopic sinus surgery. *Surgeon*. 2008;6(1):37–44.
11. Vassiliou MC, Feldman LS, Fraser SA, et al. Evaluating intraoperative laparoscopic skill: direct observation versus blinded videotaped performances. *Surg Innov*. 2007;14(3):211–216.
12. Kurashima Y, Feldman LS, Al-Sabah S, et al. A tool for training and evaluation of laparoscopic inguinal hernia repair: the global operative assessment of laparoscopic skills–groin hernia (GOALS-GH). *Am J Surg*. 2011;201(1):54–61.
13. Stack BC Jr, Siegel E, Bodenner D, et al. A study of resident proficiency with thyroid surgery: creation of a thyroid-specific tool. *Otolaryngol Head Neck Surg*. 2010;142(6):856–862.
14. Francis HW, Masood H, Chaudhry KN, et al. Objective assessment of mastoidectomy skills in the operating room. *Otolol Neurotol*. 2010;31(5):759–765.
15. Ishman SL, Brown DJ, Boss EF, et al. Development and pilot testing of an operative competency assessment tool for paediatric direct laryngoscopy and rigid bronchoscopy. *Laryngoscope*. 2010;120(11):2294–2300.
16. Andolsek K, Padmore J, Hauer KE, et al. Clinical Competency Committees: A Guidebook for Programs. Chicago, IL: ACGME; 2015. https://www.acgme.org/Portals/0/ACGMEClinicalCompetencyCommitteeGuidebook.pdf. Accessed September 5, 2017.
17. Chan T, Sherbino J; McMAP Collaborators. The McMaster Modular Assessment Program (McMAP): a theoretically grounded work-based assessment system for an emergency medicine residency program. *Acad Med*. 2015;90(7):900–905.
18. Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. *Science*. 1989;243(4899):1668–1674.

Both authors are at McMaster University, Hamilton, Ontario, Canada. **Jonathan Sherbino, MD, MEd,** is Assistant Dean, Program for Education Research and Development, and Associate Professor, Division of Emergency Medicine, Department of Medicine; and **Geoff Norman, PhD,** is Professor Emeritus, Program for Education Research and Development.

Corresponding author: Jonathan Sherbino, MD, MEd, Hamilton General Hospital, McMaster Clinic, 237 Barton Street E, Hamilton, ON L8L 2X2 Canada, 905.527.4322 ext 73542, fax 905.527.8457, sherbino@mcmaster.ca