Catching the Hawks and Doves: A Method for Identifying Extreme Examiners on Objective Structured Clinical Examinations”

July 20, 2011

Abstract

Performance-based assessments are powerful methods for assessing medical knowledge and clinical skills. Such methods involve the use of human judgment and as such are vulnerable to rater effects (e.g., halo, rater leniency/harshness, etc.) Making valid inferences from clinical performance ratings, especially for high-stakes purposes, requires monitoring and controlling rater effects. We present a simple method for detecting extreme raters in high-stakes OSCE conducted by the Medical Council of Canada (MCC). This method does not involve sophisticated statistics knowledge and can be used in any context involving human raters.

Introduction

Performance-based assessments are commonly used to assess medical knowledge and clinical skills. Objective structured clinical exams (OSCEs) are the preferred method for assessing many of these skills. An OSCE is a timed, multi-station examination in which candidates perform authentic tasks such as interviews, physical exams, and counseling with simulated or standardized patients (SPs) in realistic settings. At each station the candidate's performance is evaluated with specific checklists and rating scales by a rater who could be a physician, an SP, or some other type of observer. The whole examination is a circuit of typically 10-20 stations, with the candidates moving from one station to another until all stations are completed. At each station, the candidate is typically evaluated by a different rater.

Performance-based assessments have some clearly identified strengths and weaknesses. In terms of strengths, such assessments provide a way of observing examinees' application of knowledge and skills on authentic tasks; tasks that simulate real world tasks. Such assessments provide an opportunity to assess a wide range of competencies on both well- and ill-structured problems. Performance assessments like OSCEs are attractive because they enable the same complex and realistic clinical scenarios to be presented to many candidates in objective ways. As such, they have become the gold standard for performance-based assessment of clinical skills in medicine. The reliability and validity of scores on OSCE stations is dependent, at least to some extent, on the accuracy of the judgments.

In terms of weaknesses, scores obtained from performance-based assessments where clinical experts or SPs rate the candidates' performances (e.g., an OSCE) are vulnerable to multiple sources of measurement error, including variability due to the raters. "All assessments that depend on human raters are vulnerable to mischief due to raters. Because medical education depends so heavily on assessments using human raters, serious consideration might be given to finding useful, defensible, and legitimate statistical means to detect and reduce or minimize such mischief" [1].

Considerable research has been conducted that analyzes the variance introduced by different raters across a variety of forms of testing [2-4]. In general, variability due to examiners is considered to be highly undesirable with potential negative impacts on the validity of scoring outcomes. Essentially, such error may introduce additional construct irrelevant variance into the scores and thereby compromise the ability to make meaningful inferences from the scores.

Generalizability analysis is one method for estimating various sources of error operating in a rating situation. The literature on g-studies suggests that the error variance introduced by examiners is smaller than the task sampling variance [4] , yet the variance caused by examiners should not be ignored, especially for high stake examinations.

Furthermore, the literature suggests that the amount of variance caused by examiners can be quite substantial; for example, researchers have estimated that 12% of variance is due to differences between examiners in leniency-stringency in the British clinical examination [5] . In the USA, reports on the clinical skills exam used to assess international medical graduates (IMGs) have also shown similar variance estimates. For example, Boulet reported an associated value of 10.9%[4], Roberts [3] reported a value of 8.9%, and Harasym et al., [2]claim that in their data "Examiner variance was more than four times the examinee variance".

The phenomenon of examiners being on two ends of the scoring spectrum i.e., some of the examiners scoring very leniently and others scoring very harshly, has been recognized since the beginning of the 20th century [2], and it is often referred to as the Dove/Hawk phenomenon. Dove, used in English as a term of endearment, describes lenient examiners. Hawk, used in English as a portrayal for any form of predator, describes stringent examiners.

One of the Medical Council of Canada (MCC) examinations is an OSCE and it relies on human raters. A simple generalizability study of a few stations across three administrations of the MCC OSCE data indicated that 13-17% of station score variance in this examination is due to raters. Consequently MCC is exploring methods to detect and mitigate the dove and hawk phenomenon and the systemic measurement error these examiners may cause. Systematic

measurement error due to raters is present "when the mean of a rater summed over candidates differs from the mean of all other raters"[6].

One strategy commonly used to reduce extreme examiner scoring is through training, although this can be time consuming and expensive, and the literature on training effectiveness is not consistent. Some studies have found null or even negative effects of training [7-8] and even when training effects are positive there is also evidence that the benefits of training may dissipate over time [9]. Thus, even though some forms of training may prove to be effective in reducing rater bias; the cost of training and maintaining training effects may be prohibitive for large scale examinations.

Another strategy would be to use pairs of examiners rather than individuals to ameliorate the effects of extreme raters [10]. However, this technique may not always work. Humphries and Kaney [11] delegated two examiners per station but even this improvisation could not completely eliminate the possibility of examiner bias. Furthermore, this can double the number of examiners required, imposing budgetary demands that may not be feasible for large scale examinations.

Several statistical methods can be used to measure and correct for rater effects in a post-hoc manner (e.g., multi-faceted Rasch models, linear and non-linear regression models). The use of some of these methods requires assumptions not met by the current format of the MCC OSCE which is called the Qualifying Examination Part II (MCCQE Part II). For example, the assumption of local item independence is violated when cases are not strictly comparable, or when raters score the same students. Also, some methods may be difficult or time consuming to employ as a part of a routine quality assurance procedure. Thus the current study focuses on describing a simple and flexible method that is easy and appropriate to use in such contexts.

Method

The MCCQE Part II is a three-hour OSCE that assesses the competence of candidates; specifically assessing the knowledge, skills, and attitudes essential for medical licensure in Canada prior to entry into independent clinical practice. The MCC OSCE format is carefully

designed to provide an *objective* assessment of the candidates based on a structured approach. SPs are trained to present to and interact with the candidate in standardized ways. Examiners undergo common orientation and training to reinforce their assigned task. There are clear and firm rules for how candidates, SPs, and physician examiners interact with one another as an OSCE station usually includes an examiner, an SP and a candidate.

The MCCQE Part II is a 14 station examination that is administered three times per year. For each administration of the examination the MCC hires hundreds of physicians who rate the performance of the candidates. Since the examination runs in 17 different centers, not all examiners see the same candidates. Furthermore, the examination content is based on a blueprint that specifies the medical competencies that need to be measured during each administration of the exam, but the stations vary from one administration to another. Also, since the stations vary so do the SPs, as the SPs have to reflect the demographic criterion of the presented medical problem. Different stations also differ in difficulty. Thus, cases are not usually strictly comparable. The use of different raters and SPs for each of the administrations also introduces additional variation.

Given the variety of sources of variation in the MCC OSCE (e.g., examiner, examinee, center, administration, etc.) the application of many of the well-known methods to monitor raters' performance, such as multi-faceted Rasch models would be difficult. Thus, for the MCCQE Part II there was a need to design an alternative method that could be used from one test administration to another. Such a method was devised and is referred to as the "simple method of monitoring Hawks and Doves". The method is intended for use by anyone with a basic knowledge of statistics. Provided there is enough data available for each rater, this method can be used in other assessment situations involving human raters, including high stakes examinations and university curriculum examinations.

From very early on, two main categories of examiner errors described in the literature: *correlational* errors where raters demonstrate a tendency to evaluate an examinee holistically, without discriminating between different dimensions of behavior or performance [12]; and *distributional* errors where raters fail to make adequate use of the full range of a rating

instrument [13]. Correlational errors result when examiners rate a given candidate similarly across rating scales, ignoring variation in performance across different rating components (e.g., a halo effect). Two common distributional errors are (1) range restriction and (2) leniency/severity errors. Leniency/severity errors occur when a rater systematically rates candidates too kindly or too harshly. The simple method described here is designed to pick out distributional errors. It detects examiners who demonstrate a restricted range in their ratings and who consistently rate candidates more leniently (doves) and those who tend to rate more severely than other examiners (hawks).

Data

We examined three years of MCC OSCE data. This included six examination administrations, 9 separate forms and 2,182 physician examiners rating 3,861 examination performances. Some examiners worked on a single administration during this time frame while others worked on more than one administration. For each administration, examiners rated a maximum of two stations (one in the morning and a second different station in the afternoon). The number of candidates any given examiner rated on a station ranged from an average of 16 to 36. The scoring rubric applied by the MCC for the OSCE is an extensive checklist, designed individually for each station, with 15-40 items per case on average. The checklists are generally thought to promote more consistent grading; however, some of the assessments for individual items on the checklist can be left to the interpretation of the examiner.

The simple method to monitor Hawks and Doves

We followed a simple three step procedure to classify examiners as doves or hawks. Step one was used to identify potential extreme examiners by comparing an individual rater's mean score to the mean of all raters for that station.

In step one all examiners whose *average* score for a station was more than three standard deviations above (potential dove) or below (potential hawk) the average score for all remaining

examiners on that same station were targeted as potential extreme examiners (see figure 1). The three standard deviations were selected in order to identify the very extreme ratings. As mentioned earlier, each candidate is rated by 14 different raters so we expect that a candidate may be rated by an examiner who may be a hawk, they are also rated by other examiners who may be doves, and the difference in extremes will usually equalize in the total score across cases. With the definition of three standard deviations we targeted a very small percentage of the examiner population. However, the criterion could be changed to two or two and a half standard deviations, as the examination administrators see fit.

In step two the *distribution* of ratings from extreme examiners identified in step one were compared to the distribution for all examiners to determine whether the examiner demonstrated adequate variability in their candidate ratings for a given station. The analyses in step two have a double purpose. The first purpose is to eliminate the possibility that the stations' unique scoring key causes the extreme scoring. For example, stations with a small range of possible scores may engender more extreme scoring than stations with wide range of possible scores. The second purpose is to evaluate if the extreme rater is able to discriminate among the candidates, at least to a certain degree. The examiner might have failed everybody, or almost everybody, yet he/she may still have assigned higher scores to better candidates and lower scores to worse candidates. In this case the extreme examiner's rating is still providing valuable information regarding differences in candidate performance and such an examiner is thus eliminated form further scrutiny with regard to their scoring practices (see figure 2).

The final step in identifying extreme raters, the cohort criterion, is to determine whether the candidate cohort seen by the examiners in question demonstrated adequate variability. For example, if a large majority of candidates rated by a potential dove consistently performed higher than average on the other stations then we would not classify the examiner as a dove. Similarly if most of the examinees that a potential hawk rated were poor candidates then we might no longer classify the examiner as a hawk. Here, we use data from all stations to determine candidate ability overall and to isolate extreme judgments compared to other examiners.

Results

Identification of Doves/Hawks

Out of the 2,182 examiners (3,861 examiner performances in our sample) we identified a very small subset of raters who are extreme raters according to our criteria. Application of step one, which considers the average rating of examiners, identified 33 potential dove/hawk examiners across 44 stations. The number of stations is higher than the number of examiners because some of these examiners were identified as extreme on more than one station. From these 33 doves/hawks (44 stations), almost half (17 examiners over 21 stations) were eliminated in step two. This step was performed to separate out those examiners whose ratings, although extreme, still differentiated ratings based on a restricted range of scores, and that were very severs. In this step we insured that no examiner was labeled dove or hawk simply because their station had a very narrow score range. Of the 17 remaining examiners at the end of step two, only seven remained over 17 stations after taking step 3 where the characteristics of the cohort observed by the examiners were taken into consideration.

In sum, the application of these three simple criteria identified seven of 2,182 examiners as potential targets for examiner remediation. These criteria were very straight-forward, graphical and require minimal statistical knowledge, while being quite flexible and robust. Furthermore, they are generalizable to many situations using human raters. By comparing rater's average scores within a station, these criteria control for differences due to variation in station difficulty and other potential unique characteristics. The average scores of a potential hawk on a difficult psychiatry station is only compared to the scores of his/her peers on the same station. By examining the spread of raters' scores, this method also takes into account differences due to the station scoring key and weeds out examiners who may be extreme relative to their peers from those who are extreme and who fail to demonstrate adequate variability in their scoring. Finally, by considering the cohort characteristics of examinees, this method controls for differences due to candidate abilities.

The dove and hawk analyses can be a useful part of the quality assurance used routinely at the Medical Council of Canada, as, "it is clear that standardized patients assessments can be modified and improved by systematically investigating, analyzing, and tracking candidates' scores" [4]. The performance of our examiners is monitored for each administration of the MCC OSCE, using this procedure in addition to other operational procedures that standardize the OSCE assessment program.

Discussion

While little is known about the success of mitigation strategies, basic measurement principles dictate that it is important to monitor this potential source of error and address issues as these are identified. Given that performance ratings are often used to make high-stakes decisions about an individual's career, monitoring the frequency and sources of this type of error is an important quality control mechanism to ensure that scores have the expected comparability and the pass or fail decision is valid.

The dove and hawk analysis is one of many quality assurance steps that can be taken in the process of reducing the rater error. Although the design of the OSCE program aims to offset examiner variances, we are aware that for the MCC OSCE 13-17% of variance in stations scores is due to raters. This means the extreme examiner phenomenon deserved continued attention in our research and operational QA procedures.

In this paper we have focused on the systematic error defined by the aberrant mean score of the examiners. In the process of improving examiner performance and monitoring rater bias, our future research will be extended to examining rater bias of centrality, i.e., identifying examiners who score all candidates around the borderline. A lack of discrimination among candidates lowers the reliability of total scores and suggests scales or training are inadequate. One might presume that examiners need a clearer understanding of anchor terms or may have a tendency to avoid classifications where the stakes are high for examinees. Either issue suggests score interpretations may lack validity.

Conclusion

Complex large scale OSCEs could benefit with a straightforward method for identifying extreme raters. Recognizing restricted ranges in scoring, potential differences due to station difficulty, and cohort characteristics of candidates are relevant factors in this analysis. The proposed computations may help to identify extreme raters who may raise concerns, and can be used to improve the comparability or scores and the confidence we have in our decisions using OSCEs.

Having a clear set of simple procedures in place for identifying extreme raters has both practical and theoretical significance. From a practical standpoint, this is a useful QA step to identify doves and hawks. Examiners identified by these steps can be contacted and their scoring discussed to align it with the scorings of their peers. Additional training for some examiners may be offered in the future. Examiner performances will be tracked in a data base. An examiner who is repeatedly classified as a dove or hawk and who does not appear to alter their scoring with feedback and training will be removed from the MCC examiner pool. The tracking of examiner performances can be a critical source of quality management when performances are complex and rely on human judgment.

References

1. Downing, SM. Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education* 2005; *39*: 350-355.

2. Harasym, PH, Woloschuck, W, Cunning, L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Advances in Health Sciences Education* 2008; *13*: 617-632.

3. Roberts, C, Rothnie, I, Zoanetti, N, et al. Should candidate scores be adjusted for interviewer stringency or leniency in multiple mini-interviews? *Medical Education* 2010; *44*: 690-698.

4. Boulet JR, Mckinley DW, Whelan GP, et al. Quality Assurance Methods for Performance-Based Assessment, *Advances in Health Sciences Education* 2003; *8: 27-47.*

5. McManus, IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES): using multi-facet Rasch modeling. *BMC Medical Education.* 2006;6:42. doi: 10.1186/1472-6920-6-42.

6. Raymond, MR, Viswesvaran, C. Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement* 1993; *30*: 253-268.

7. Bernardin, HJ, & Buckley, MR. Strategies in rater training. *Academy of Management* 1981; *6*: 205-222.

8. Iramaneerat C, Yudkowsky R. Rater errors in a clinical skills assessment of medical students. *Evaluation of Health Professionals* 2007;*30* 266-83.

9. Ivancevich, JM. Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology* 1979; *64*: 502-508.

10. Muzzin, LJ, Hart, L. Oral examinations. In: Neufeld, Victor R, Norman GR (eds*)*, *Assessing Clinical Competence*. Springer, New York. 71-93. 1985.

11. Humphris, G, Kaney, S. Examiner fatigue in communication skills of objective structured clinical examinations. *Medical Education* 2001; *35*: 444-449.

12. Thorndike, EL. A constant error on psychological rating. *Journal of Applied Psychology 1920; IV*: 25-29.

13. Kingsbury, FA. Analyzing ratings and training raters. *Journal of Personnel Research* 1922; 1: 377-388.

Figure 1. Step one: average station score for examiners are compared to the average of all remaining examiners to identify potential hawks (3 SD below the distribution of all examiners) and doves (3 SD above the distribution of all examiners).

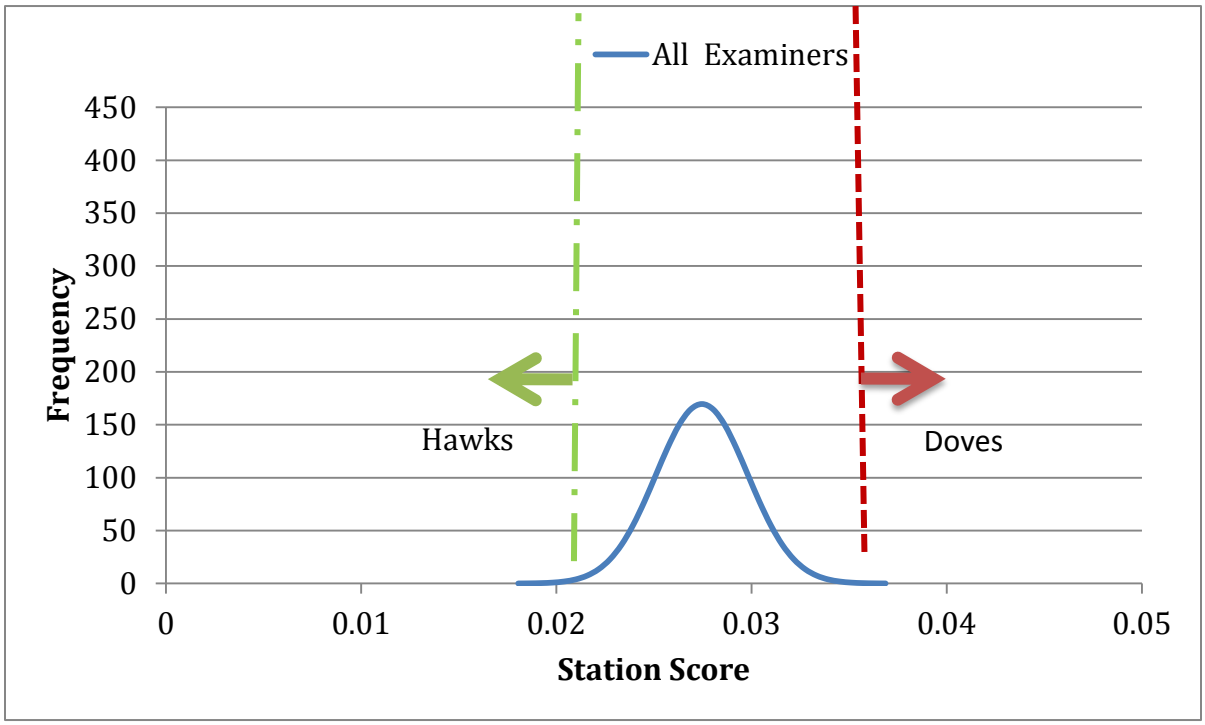

Figure 2. Step two: the distribution of examiners identified in step one are compared to the average of all remaining examiners to check for variability of ratings.