Education in Medicine Journal ISSN 2180-1932

INNOVATIVE IDEAS
Volume 4 Issue 2 2012
DOI:10.5959/eimj.v4i2.105
www.eduimed.com

The discrepancy-agreement grade (DAG): a novel grading system to provide feedback on rater judgments.

Muhamad Saiful Bahri Yusoff, Ahmad Fuad Abdul Rahim

Medical Education Department, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kota Bharu, Kelantan, Malaysia

ARTICLE INFO

Received : 25/07/2012
Accepted : 27/09/2012
Published : 01/12/2012

KEYWORD

Inter-rater reliability
Rater judgment
Intra-class correlation
Assessment
Interview
Examination
Agreement level

ABSTRACT

Introduction: Ratings are known to have a generosity error, provide limited discrimination and distorted interpretation, and often fail to document serious deficits. A potential source of these problems is rater judgement. These problems compromise the capability of raters to maintain the standards of rating. The authors propose a simple grading system to improve this situation including providing feedback to raters. **Method:** The authors developed a grading system named the Discrepancy-Agreement Grade (DAG) to provide feedback on rater judgments. Dependent-t and intraclass correlation tests were applied to determine discrepancy and agreement levels of raters. Rater judgments were then classified into grades A, B, C or D. This grading system was tested in an examination and a student selection interview to assess rating judgments of examiners and interviewers. The purpose was to evaluate the practicability of the grading system to provide feedback on examiners' and interviewers' rating judgements. **Results:** in the examination, five short essays were rated by five pairs of senior lecturers. Out of 5 pairs, 2 (40%) obtained grade A and 3 (60%) obtained grade B. In the student selection interview, a total of 48 pairs of interviewers interviewed ten applicants. Out of 48 pairs, 20 (41.7%) obtained grade A, 1 (2.1%) obtained grade B, 23 (47.9%) obtained grade C and 4 (8.3%) obtained grade D. **Conclusion:** The grading system showed variability of rater judgments on medical students' and applicants' performance in an examination and interview session respectively. It provided feedback on the examiners' and interviewers' judgments on candidate performances. This exercise demonstrated practicability of the grading system to provide feedback on rater judgements.

CORRESPONDING AUTHOR: Dr Muhamad Saiful Bahri Yusoff, Medical Education Department, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kota Bharu, Kelantan, Malaysia.
Email: msaiful@kb.usm.my

Introduction

Changes in the education environment are putting pressure on medical schools to ensure faculty accountability on assessment decisions made and to document the quality of medical education they provide (1). Faculty judgments on medical students' performance during medical training are an important element in this effort to document such quality procedures.

Faculty judgments are known to have a generosity bias, often fail to document serious deficits, are prone to distorted interpretation and often provide limited discrimination on students' performance (1). The potential sources of judgment errors are the individual as the rater, the form or scales used for rating, the items used for rating and the objects of rating (2). These problems compromise the capability of raters to maintain the standards of rating.

Based on the literature, it appears that the main source of rating error is the faculty judgment variability, typically ranging between 80% to 90% (2). Therefore, relevant feedback regarding faculty rating performance will help to improve their rating performance in the future. As reported by previous studies, feedback is a powerful tool to improve individual performance (3-6). The feedback will provide valuable information to the faculty on their judgement on the student performance. Feedback produces the most impact when goals of feedback are specific, is done in a non-threatening environment and builds on changes from previous experience (3-6). Therefore, it is timely that the authors introduce a grading system as a mechanism to provide relevant feedback to the faculty about their rating performance.

The authors designed two studies to evaluate the practicality of the grading system in providing feedback on the faculty rating performance during examination and interview situations. The authors hypothesised that if the grading system is practical and valid, it will show lesser faculty rating variability during the examination compared to the interview session. This is due to the examination being more structured than the interview session, thus leading to less faculty rating variability during the examination

Method

The Discrepancy-Agreement Grade (DAG)

The DAG was developed to measure inter-rater variability based on two statistical tests which are the dependent-t and intraclass correlation (ICC). The dependent-t test and ICC are applied to determine discrepancy and agreement between two raters respectively. The discrepancy is considered non-significant if p-value of the dependent-t test is more than 0.05. The agreement level is considered as acceptable when the ICC value more than 0.4 (7-10). Based on the results of the two tests, rater judgments are classified into grades A, B, C or D (See Table 1).

Grade A is the best condition where the two raters have a good agreement level and are scoring with similar weightages whereas Grade D is the worst condition where the two raters are in poor agreement and scoring with dissimilar weightages. Grade B is the condition where the two raters are scoring with different weightages but have a good agreement level therefore taking mean marks of both raters is recommended. Grade C is the condition where the two raters have a poor agreement level but with no obvious discrepancy of mean marks given. Remarking after discussion is recommended for grades C and D. Grade A and B are considered as a good level while grade C and D are considered as a poor level of rater judgments. The description of each category was summarised in the table 1.

Study design and sample

Two cross-sectional studies were conducted on two occasions which involved 1) five pairs of senior lecturers who rated five short essay questions (SEQ) in a final examination and 2) 48 pairs of experienced interviewers who interviewed applicants of a medical degree program. The discrepancy and agreement level

between two examiners for each pair were analysed and graded based on the DAG grid (table 1). Each examiner and interviewer was given an identity code to ensure they are anonymous and for follow up purposes. Each pair of SEQ examiners rated answer sheets of 44 medical students and each pair of interviewers interviewed 10 applicants. Permission to conduct the studies was obtained from the medical school prior to the start of the studies.

Data collection

SEQ marks given by each examiner and interview marks given by each interviewer were obtained from the academic office.

Statistical analysis

SPSS version 18 was used to analyse the data. ICC analysis was performed to measure inter-examiner reliability and dependent-t test was performed to measure discrepancy between the two examiners. Data was cleaned prior to the analyses.

Table 1: The Discrepancy-Agreement Grade (DAG) Grid: possible results of data analysis and recommended actions

		Discrepancy (dependent-t test)			
		Not significant (p > 0.05)		**Significant (p < 0.05)**	
Agreement Level(ICC)	**Acceptable Agreement (1> ICC > 0.4)**	**++**	**Grade A** **The best situation.** **Raters are in agreement and are scoring students with similar weightages.** **Recommended action: Marks scored by the examiners can be used.**	**+-**	**Grade B** **The second best situation.** **Raters are in agreement but are scoring students using different weightages.** **Recommended action: Determine if raters are giving more or less marks; discuss possible reasons. Recommend either re-marking done after discussion or take the mean marks of raters.**
	Non-acceptable Agreement (ICC < 0.4)	**-+**	**Grade C** **The second worse situation** **Raters are not in agreement but the mean of marks given is not significantly different.** **Recommended action: Re-marking after discussion.**	**--**	**Grade D** **The worst situation** **1) Raters are not in agreement and the mean of marks given are different.** **Or** **2) raters are in perfect agreement (ICC = 1) and the mean of marks given are the same (i.e. an indication for the raters are not rating independently).**[1] **Recommended action: Re-marking after discussion.**

Results

Five short essays were rated by five pairs of senior lecturers. Out of 5 pairs, 2 (40%) obtained grade A and 3 (60%) obtained grade B. These results showed that there was good agreement level between the examiners and less discrepancy.

A total of 48 pairs of interviewers interviewed ten applicants. Out of 48 pairs, 20 (41.7%) obtained grade A, 1 (2.1%) obtained grade B, 23 (47.9%) obtained grade C and 4 (8.3%) obtained grade D. These results showed that there was variability between the interviewers with regards to discrepancy and agreement levels. These results supported the hypothesis that the grading system is practical and valid; it showed lesser faculty rating variability in the examination than the interview session.

Discussion

Rater judgments are known to have a generosity error, provide limited discrimination and distorted interpretation and often fail to document serious deficits (1). The potential sources of these problems are related to the mechanics of the rating task, the system used to obtain ratings and factors affecting rater judgement (1). These problems compromise the capability of raters to maintain the standards of rating. The authors call for an effort to stop the erosion of standards by a simple grading system.

The results showed a low level of variability among examiners in an examination and a high level of variability between interviewers in rating the performance of candidates. These results supported the practicality and validity of the grading system to provide feedback on rater judgements.

The DAG was able to provide feedback with a specific direction (i.e. agreement and discrepancy levels). In addition, information was obtained in a non-threatening manner. It also provided specific information based on previous rating performance. All these are good characteristics of a feedback mechanism (3-6). These findings suggest that the DAG might serve as a promising feedback tool to improve rating performance among faculty in educational institutions. However, many further research needs to be done to provide evidence of its practicality and usefulness in various contexts.

It is worth highlighting that there are many ways to minimise influences of the potential sources of rating errors. To minimise the rating errors related to the raters, perhaps a few approaches can be introduced which include faculty training to observe the quality or attributes being rated, familiarising faculty to the rating system being used to rate student performance and establishing clear expectations on rater roles in rating performance (2). To minimise the rating errors related to the scales and items medical schools should standardize the rating form and make the rating task as easy as possible (2). To minimise the rating errors related to the objects being rated, inter-rater reliability needs to be improved and a triangulation technique can be used whereby a decision on student performance are decided based on multiple assessment tools (2). It is worth stressing that most of the sources of rating error are modifiable if appropriate feedback given on rating performance to the faculty. This is where the DAG can play a role in providing such feedback to the raters.

The advantages of using the DAG include the simplicity of the grading system and ease of application, yet at the same time able to give specific information on the rating performance of raters. In addition, it is done in a non-threatening way. Therefore, the DAG system has tremendous potential to be adopted by medical and allied health schools as a feedback mechanism to rater judgments in various educational settings.

Conclusion

The grading system showed variability of rater judgments on medical students' and applicants' performance in an examination and interview session respectively. It provided feedback on the examiners' and interviewers' judgments on candidate performances. This exercise demonstrated practicability of the grading system to provide feedback on rater judgements.

Conflict of interest

Part of this article was published in Medical Education Journal under the Really Good Stuff: Yusoff MSB. Discrepancy-agreement grading provides feedback to rater judgments. *Medical Education*, 2012; 46(11): 1122.

Reference

1. Albanese MA. Challenges in using rater judgements in medical education. Journal of evaluation in clinical practice. 2000;6(3):305-19.
2. Downing SM. Threats to the validity of clinical teaching assessments: What about rater error? Medical education. 2005;39(4):353-5.
3. Hattie J. Influences on student learning. Inaugural lecture given on August [serial on the Internet]. 1999; 2: Available from: http://www.education.auckland.ac.nz/webdav/site/education/shared/hattie/docs/influences-on-student-learning.pdf.
4. Hattie J, Timperley H. The power of feedback. Review of educational research. 2007;77(1):81-112.
5. Kluger AN, DeNisi A. Feedback interventions: Toward the understanding of a double-edged sword. Current Directions in Psychological Science. 1998;7(3):67-72.
6. Norcini J. The power of feedback. Medical Education. 2010;44(1):16-7.
7. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977:159-74.
9. Shrout PE, Fleis JF. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin. 1979;86(2):420-8.
10. Streiner LD, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. 4th ed. New York: Oxford University Press; 2008.