Education 2014, 4(1): 12-17
DOI: 10.5923/j.edu.20140401.03

Inter-rater Reliability and Agreement of Rubrics for Assessment of Scientific Writing

Eva Ekvall Hansson*, Peter J. Svensson, Eva Lena Strandberg, Margareta Troein, Anders Beckman

Lund University, Department of Clinical Sciences in Malmö/Family Medicine

Abstract *Background:* Learning how to write a scientific paper is a mandatory part of medical education at many universities. The criteria for passing the exam are not always clear; the grading guidelines are often sparse and sometimes poorly defined. Therefore, the use of rubrics can be appropriate. *Purpose:* The aim of this study was to test inter-rater reliability and to test agreement for the modified rubrics for the assessment of master's theses in medical education at a Swedish university. *Method:* Modified scoring rubrics were used for grading and assessment of the master's thesis at the medical programme at Lund University. The rubrics include 10 items, graded from 1 to 4. To study the inter-rater reliability and agreement of the rubrics, three teachers included in the management of the course used the rubrics and assessed all projects. *Results* A total of 37 projects were read by the three raters. Intraclass correlation for the total score was 0.76 (CI 0.59–0.87). Absolute agreement (average) for pass or fail was 90%. *Conclusion:* In this study, scoring rubrics for assessing master's theses in medical education showed strong inter-rater reliability and high inter-rater agreement for pass/fail. The rubrics are now available on the university website.

Keywords Scoring rubrics, Inter-rater reliability, Inter-rater agreement, Master's theses

1. Introduction

Learning how to perform a scientific study and write a paper is a mandatory part of medical education at many universities. In Europe, work has been done to adjust medical education according to the Bologna process[1] and higher education in the whole health care sector will probably benefit from these efforts[2]. On the medical education programme at Lund University in Sweden, one result of this work is that the course in scientific writing has been changed from a 10-week course to a 20-week course at master's level[3]. The Swedish National Agency for Higher Education is the public authority that reviews the quality of higher education institutions in Sweden. The agency regards the scientific projects produced in higher education as important[4]. Considering this, teachers on the medical programme have to relate to whether the scientific project should be examined as the process of learning how to write a scientific project or as the actual production of a paper[5]. The courses comprise work-based learning, meaning that the students, under supervision, perform a study and write a scientific paper. The courses are usually examined in a threefold way: the production of a scientific paper, an oral presentation and defence of the paper, and opposition on another student's paper. Hence, it appears that the production of the paper is examined, not the process. Also, the criteria for passing the exam are not always clear; the grading guidelines are often sparse and sometimes poorly defined[6]. At the start of a course, students should be informed about criteria and how to pass the exam, but this is not always clear when dealing with scientific projects. Other persons also need to be fully aware of the criteria: the student, the supervisor, the examiner and the head of the course.

1.1. Constructive Alignment

Long-term educational goals are needed to enhance advanced knowledge in higher education[7]. Constructivist learning theory can help teachers in higher education with this enhancement[8]. Constructive alignment has been suggested as a way to put together two of the important lines of thinking about teaching and learning in higher education: constructivism and instructional alignment[8]. The concept is student-centred and outcome-focused and intended learning outcomes, learning activities and assessment tasks are in alignment to each other[9]. Constructive alignment can well be used in the context of work-based learning[10] such as medical education programmes.

1.2. Rubrics

To facilitate constructive alignment and make the alignment between learning outcomes, learning activities and assessment task more evident, the use of rubrics can be appropriate[11]. Rubrics can also be relevant in order to

* Corresponding author:
eva.ekvall-hansson@med.lu.se (Eva Ekvall Hansson)

Published online at http://journal.sapub.org/edu

avoid arbitrariness and to stimulate learning[11]. Expectations and criteria are made explicit when rubrics are used[12]. Providing examiners with detailed rubrics can improve the quality of the examined task and the generalizability of the rubrics used[13]. Also, for the correctness of the outcome in research on assessment, carefully designed instruments are important[14]. It is essential to communicate learning outcomes and make them evident to the students, and rubrics can help teachers in higher education with this achievement[7].

Clarity and appropriateness of language seem to be central concerns when using rubrics for grading in higher education [11]. The usefulness of an assessment tool is determined by its standard in fulfilling accepted criteria, i.e. to be reliable, valid, feasible, fair and beneficial to learning[15].

1.3. Reliability

The consistency of the assessment instrument when it is repeated is referred to as reliability[14]. Validity is the extent to which an assessment measures what it is supposed to measure[16], and more attention to validity and reliability is suggested[11]. Reliability also refers to the generalizability of the assessment measure and reliability coefficients concern the estimation of random errors of measurements in assessment, thus improving the overall assessment[17]. Inter-rater agreement is the extent to which assessors make exactly the same judgement about a subject[18]. Since the interpretation and synthesis of study results are often difficult, guidelines for reporting reliability and agreement studies have recently been proposed[19].

In 2010, scoring rubrics for grading and assessment of master's theses that use quantitative methodology were developed at Lund University[20]. These scoring rubrics have been modified for use for quantitative as well as qualitative methodology by the authors of the present paper.

We had a twofold aim in this study: to test inter-rater reliability and to test agreement on the modified rubrics for the assessment of master's theses in medical education at a Swedish university.

2. Methods

2.1. The Examination

At the medical school of Lund University, the course in scientific writing was placed in the eleventh and last semester, extending over ten weeks, when this study was performed. The course comprised the writing and presentation of a scientific paper. The course is examined in a threefold way: the production of a scientific paper, an oral presentation and defence of the paper and opposition on another student's paper. Together, this threefold examination comprises the final examination which is non-graded (pass/fail), where the achievements of the intended learning outcomes for the course are assessed. The head of the course was responsible for the whole examination, which included the written paper, the oral presentation and the task as opponent. In order to ensure sufficient quality, the students had to deliver their scientific paper four weeks before the oral presentation and the paper had to pass through a pre-exam inspection. The pre-exam inspection was carried out by the head of the course.

The students made an oral presentation of their paper at a seminar. At the seminar, a fellow student acted as primary opponent, and an external examiner, who was an expert in the research field in question, acted as secondary opponent and had also evaluated the project before the seminar. At the same seminar, the student also acted as opponent on a fellow student's project.

2.2. Inter-reliability and Agreement of the Rubrics

Three teachers (EEH, PJS, MT) included in the management of the course made a preliminary assessment of all projects registered for presentation at the seminars at Campus Malmö during the spring of 2011. These three teachers were experienced in reviewing scientific work from the medical students. The rubrics were used to grade the papers. After this, the preliminary grading of the papers was discussed on a general level at a seminar where the three teachers analysed the assessment process and how the scoring rubrics were used, in order to reach agreement. After this seminar, the teachers made a final grading.

The development of the rubrics was based on prior work in this area: the rubrics developed by Jernström[20) were originally created to assess degree projects based on quantitative methodology at the master's level. They were based on the European Credit Transfer and Accumulation System scale (ECTS). The grading scale was criterion-referenced in six levels (A to F scale) with 15 different items. Six levels and 15 items were considered inappropriate to use in this context; they were therefore modified to a 1-to-4 scale and 10 items. The instructions on each item and each grade were also modified in order to make the rubrics suitable to use for both qualitative and quantitative projects. Also, the different items were not regarded as equally important for grading overall achievements. Therefore, the items are weighted. Five items were given the weight of the grade multiplied by one and five items the weight of the grade multiplied by two. The multiplication was done after the individual teachers' grading for analytic purposes. The rubrics are shown in table 1.

The scoring rubrics developed and tested in this study only cover the assessment of the thesis as a written product.

2.3. Statistics

The results of the grading were collected and analysed anonymously with respect to both students and teachers. For the analysis of agreement for pass/fail, all gradings were dichotomously transformed.

For inter-rater reliability (IRR) we analysed the intraclass coefficient (ICC), mixed model for consistency[21, 22]. ICC ranges from 0 (no agreement) to 1 (perfect agreement)[21]. Analyses were performed separately for individual rubrics, as well as for the sum of total points. The inter-rater agreement (IRA) for pass-fail was analysed using percentage agreement. IRA was analysed separately for individual rubrics, the percentage agreement for each student was calculated and the overall mean was determined[23] .

Data from the assessors' ratings of the projects were analysed using SPSS 20.0 (SPSS Inc, software location Lund University). Percentage agreement was calculated by hand.

3. Results

A total of 37 projects were assessed individually by the three raters. The projects received a mean score of 35.1 (SD 4.1), range of score 23–51, for all three raters. The IRR for individual rubrics expressed by ICC (consistency) was 0.76 (95% CI 0.59–0.87) and ICC varied from 0.15 (abstract) to 0.87 (ethics). The IRA for individual rubrics ranged from 73% (ethics) to 100% (introduction, results and references). The average for absolute agreement for total pass or fail was 90% (individual *student* range 70–100%). A comparison between each rubric and total agreement on pass-fail and ICC is shown in table 2 and figure 1.

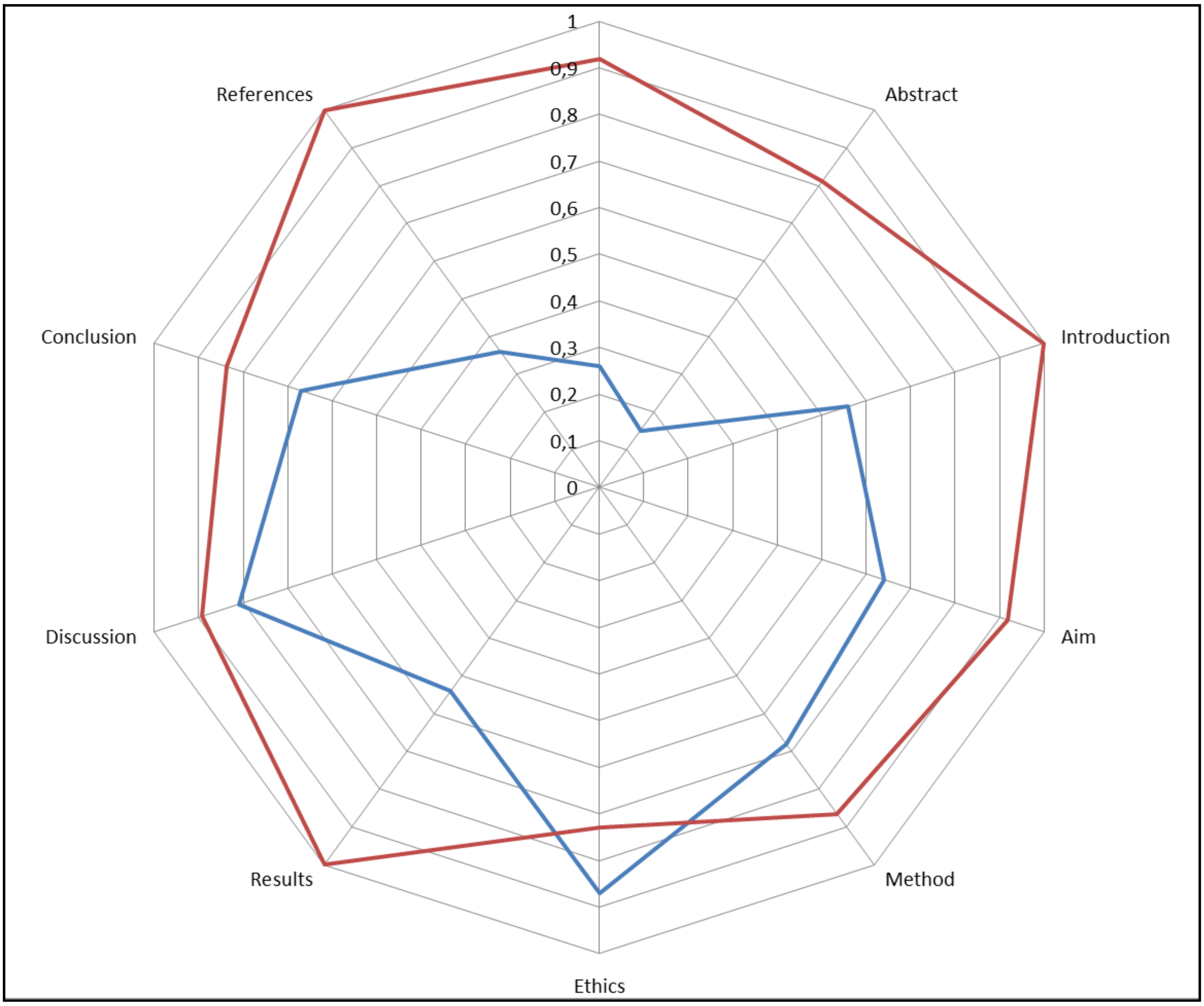

Figure 1. Coefficients for percentage agreement (red) and ICC consistency (blue)

Table 1. Rubrics used in the study. Rubrics are listed in the first column. Columns 2–4 state examples of grades. Column 5 states the weight applied to each grade

	Not sufficient **(1)**	**Satisfactory** **(2)**	**Good** **(3)**	**Excellent** **(4)**	**Weight** **W x grade**
Title	Unclear or poorly stated	Satisfactory stated title includes some keywords found in aim.	Complete title, includes keywords found in aim.	Complete title, includes clear keywords found in aim and mentioned in introduction.	W = 1
Abstract	Missing or too long or too short. Does not contain background, aim, results or conclusion	Contains background, aim, results and conclusion. The different parts of the abstract may not be in proportion to each other.	Well written with relevant background, aim, results and conclusion, in proportion to each other.	Excellently written, clear and concise abstract with highly relevant background, aim, results and conclusion, in proportion to each other.	W = 1
Introduction/ background	Missing, too short or little or no relevance to the topic. No references or inadequate references.	Gives satisfactory perspective on the problem. Gives explanation of aim. Uses adequate references.	Gives a good perspective on the problem on a national and international level. Gives good explanation of why the aim is important to study. Uses relevant references.	Gives a clear perspective on the problem on a national and international level. What is known and what is unknown. Clearly states why the aim is important to study. Uses highly relevant references.	W = 2
Aim	Missing or not relevant to the topic or inadequately formulated.	Satisfactorily formulated aim relevant to the topic	Well-stated aim, relevant to the topic.	Clearly stated aim, highly relevant to the topic. Well suited for the course.	W = 2
Methods/ Material	Not suited or poorly suited for testing the hypothesis or not described or poorly described. Not applicable or poorly used method. No references or inadequate references. Descriptions of statistical methods missing or only mentioning which program was used, not which analyses. For qualitative projects: method poorly described, analysis poorly described or analysis not applicable for method.	Adequately suited and applicable for testing the hypothesis and adequately described with adequate references. Method well applied. Statistical program stated, statistical methods used are mentioned and correct. For qualitative projects: Method and analysis adequately described. Analysis applicable for method.	Well suited and applicable. Well described with appropriate references. Appropriate and well applied method for testing the hypothesis. Statistical program stated. Statistical methods used are well described and correct. For qualitative projects: Method and analysis well described. Analysis well applied for the method.	Well suited and clearly described with correct references. Correct and very well applied method for testing the hypothesis in the chosen material. Clearly applicable. Statistical program and version stated and referenced. Correct statistical method used and clearly described with references. For qualitative projects: Method and analysis well suited and correctly presented. Analysis very well applied for the method.	W = 2
Ethics	Missing or considerations not relevant to the project.	Ethical considerations relevant to the project are described	Ethical approval from regional ethics board if appropriate. Ethical considerations, relevant to the project are well described.	Ethical approval from regional ethics board if appropriate. Current and future ethical implications of the current study are mentioned in the background and discussed in the discussion	W = 1
Results	Results not satisfactorily related to aim or mostly missing. Tables or figures with faults or missing. For qualitative projects: Quotes missing.	Too many or too few results presented, not clear which are the main ones. Tables or figures contains relevant characteristics of the study. For qualitative projects: Too many or too few quotes presented.	Results based on the aim are presented. Adequately structured. Tables and figures contains relevant characteristics of the study. Main results highlighted in table or figure. For qualitative projects: Quotes adequately presented	Main results based on the aim are clearly presented. Clearly structured. Tables and figures contains relevant characteristics of the study. Main results highlighted in table or figure. Missing data clearly indicated. Flow chart presented when appropriate. For qualitative projects: Quotes clearly presented with a good structure.	W = 2
Discussion	Not relevant to the study or poorly structured. No connection to the results. No connection to other research. Strength and weaknesses of the study are not mentioned.	Satisfactory discussion of results and the strength and weaknesses of the study	Well written discussion of main results in relation to aim. Most of the study's strength and weaknesses are discussed. Well structured.	Very well written discussion of main results in relation to aim. Strength and weaknesses and potential bias are discussed and put into a new perspective. Excellent structure.	W = 2
Conclusion	Missing or not satisfactorily related to results or not relevant to aim	States some findings but not all or not main findings	States main findings. Well formulated	State main findings and its implications in a short- and long-term perspective. Very well formulated.	W = 1
References (as a whole)	Missing, irrelevant, too basic or poorly organized references.	Satisfactory use of mostly adequate references in the correct order	Good use of relevant references properly referred to in the text. All references in the correct order and the same format.	Clearly relevant references used, skilfully referred to in the text. All references in the correct order and the same format.	W = 1

Table 2. Level of agreement for pass or fail and intraclass correlations (ICC consistency) for level of agreement and 95% confidence interval (CI) for ICC

	Agreement pass-fail (%)	ICC	95% CI
Title	91	0.26	–0.28 – 0.59
Abstract	81	0.15	–0.46 – 0.53
Introduction	100	0.56	0.25 – 0.76
Aim	92	0.64	0.39 – 0.80
Method	86	0.68	0.44 – 0.82
Ethics	73	0.87	0.78 – 0.93
Results	100	0.54	0.21 – 0.75
Discussion	89	0.81	0.68 – 0.90
Conclusion	84	0.67	0.44 – 0.82
References	100	0.36	–0.10 – 0.65
Total score, average	90	0.76	0.59–0.87

4. Discussion

In this study of inter-rater reliability and absolute agreement of scoring rubrics, the total weighted score had a strong inter-rater reliability (ICC 0.76), and the average level of absolute agreement was high (90%). For individual rubrics the inter-rater reliability varied from 0.15 to 0.81 and absolute agreement from 73% to 100%.

The item “abstract” had the lowest ICC value (0.15). One explanation might be that the assessors are experts in different research fields and the assessment of an abstract is therefore influenced by the assessor’s framework. There is a possibility that the scoring of the item “abstract” will have a higher reliability when the assessor is an expert in the research field in question.

In scientific writing, attention to sentence structure, style, and logical flow is proposed[24], an issue not taken into account in the scoring rubrics studied.

This study is small, which has to be taken into account when considering the results. However, the 37 projects included in the study are authentic and the three raters are experienced teachers in the course and we therefore believe that the results are valid.

Since Fleiss kappa is not appropriate to use when analysing dichotomized data with an uneven distribution, the ICC was calculated for measuring inter-rater reliability[25]. The variability between high percentage agreement and low ICC illustrates well the inappropriateness of analysing dichotomies with small differences using methods that assume high variability[26].

The rubrics are now available on the Lund University website and are therefore available for students, examiners and supervisors alike[3]. The examiners’ judgement about the quality of the projects is thereby undisguised for the students[27]. The instructions on the website about how to use the rubrics declare that no project can pass if any item has the grade “not sufficient” and the difference between grades 1 (fail) and 2 (pass) is therefore essential. In our study, the average level of absolute agreement was high (90%), which indicates that the students get a fair examination when the rubrics are used. Pass/fail has crucial effects for the student, and that is why the percentage agreement must be high[8]. The seminar where the teachers discussed the grading of projects on a general level seemed crucial for reaching agreement. Also, the seminar probably provided the assessors with an opportunity to achieve clarity and appropriateness in language, which is crucial for the validity of rubrics[11]. Therefore, in order to reach high agreement between assessors, we strongly recommend a discussion on a general level about assessment and the scoring rubrics before using them.

The Swedish National Agency for Higher Education is the public authority that reviews the quality of higher education institutions in Sweden. The agency regards the scientific projects produced in higher education as important[4]. Therefore, it is also important that these projects reflect the quality of the educational programme. Scoring rubrics might facilitate higher quality in scientific projects. Hence, in order to find out whether the use of rubrics has had any effect on the quality of the papers, we plan to compare projects produced before the introduction of the rubrics with projects produced after.

5. Conclusions

In this study, scoring rubrics for assessing master’s theses in medical education showed strong inter-rater reliability and high inter-rater agreement. To reach agreement, we recommend teachers to discuss the rubrics on a general level before using them.

REFERENCES

[1] European Commission. Education&Training. Available at: http://ec.europa.eu/education/higher-education/bologna_en.htm2011 [cited 2013-06-19].

[2] Hensen P. The “Bologna Process” in European Higher Education: Impact of Bachelor's and Master's Degrees on German Medical Education. Teach Learn Med. 2010;22(2): 142-7.

[3] Lund University. Faculty of Medicine, Lund 2013 [cited 2013-06-19]. Available from: http://www.med.lu.se/laekarutbildning/om_laekarutbildningen/utbildnings_och_kursplaner.

[4] Swedish National Agency for Higher Education [website]. Stockholm 2011 [cited 2011-12-15]. Available from: http://www.vhs.se/sv/In-English/.

[5] Aspegren K, Danielsen N, Edgren G. Pedagogic in Medical Education - a manual for teachers in medical education (In Swedish). Lund: Studentlitteratur; 2012. 204 p.

[6] Truemper CM. Using scoring rubrics to facilitate assessment and evaluation of graduate-level nursing students. J Nurs

Educ. 2004;43(12):562-4.

[7] Cole N. Conceptions of educational achievement. Educational Researcher. 1990;19(3):2-7.

[8] Biggs J. Enhancing teaching through constructive alignment. High Educ. 1996;32(3):347-64.

[9] Biggs J, Tang C. Teaching for Quality Learning at University. 3rd ed. Open University Press; 2007. 335 p.

[10] Walsh A. An exploration of Bigg's constructive alignment in the context of work-based learning. Assessment&Evaluation in Higher Education. 2007;32(1):79-87.

[11] Reddy YM, Andrade H. A review of rubric use in higher education. Assessment & Evaluation in Higher Education. 2010;35(4):435-48.

[12] Jonsson A, Svingby G. The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review. 2007;2:130-44.

[13] Baldwin SG, Harik P, Keller LA, Clauser BE, Baldwin P, Rebbecchi TA. Assessing the impact of modifications to the documentation component's scoring rubric and rater training on USMLE integrated clinical encounter scores. Acad Med. 2009;84 (10 Suppl): S97-100.

[14] Schuwirth L, Colliver J, Gruppen L, Kreiter C, Mennin S, Onishi H, et al. Research in assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):224-33.

[15] Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206-14.

[16] McAleer S. Choosing assessment instruments. In: Dent J, Harden R, editors. A practical guide for medical teachers. 3rd ed. Edinburgh: Churchill Livingstone/Elsevier; 2009. p. 318-24.

[17] Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38(9):1006-12.

[18] Tinsley H, Weiss D. Interrater reliability and agreement of subjective judgments. J Counseling Psychology. 1975; 22(4): 358-76.

[19] Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud. 2011;48(6):661-71.

[20] Jernström H, . Development of a scoring rubric for A to F grading and assessment of Master's Thesis that are using quantitative methodology. Paper in course in Higher Education, Lund University. Lund 2010.

[21] Kirkwood B, Sterne J. Essential medical statistics Oxford: Blackwell Publishing Company; 2003. 499 p.

[22] LeBreton JM. Answers to 20 questions about interrater reliability and interrater agreement. Organizational Research Methods. 2008;11(4):815-52.

[23] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82.

[24] Menger F, Rizvi S. An Essay on Scientific Writing. Education. 2013;3(2):130-3.

[25] Freelon D. or: American University School of Communication; 2009 [cited 2013 2013-05-02]. Available from: http://dfreelon.org/2009/12/09/from-the-mailbag-120909/.

[26] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990; 43(6): 543-9.

[27] Sadler DR. Interpretations of criteria-based assessment and grading in higher education. Assessment & Evaluation in Higher Education. 2005; 30(2):175-94.