International Journal of Linguistics and Translation Studies
Volume 1, Issue 1, 2020
Homepage http://ijlts.org/index.php/ijlts/index

Towards Consistent and Fair Assessment Practice of Students' Subjective Writing

Takad Ahmed Chowdhury
University of Asia Pacific (UAP), Bangladesh
takad@uap-bd.edu

ARTICLE HISTORY

Received: 01/02/2020

Accepted: 28/04/2020

KEYWORDS
fair assessment, rubrics, reliability, consistency, holistic marking, analytic marking

Abstract
The question of fairness is perhaps as old as the invention of assessment in education, and it is of utmost significance for the teachers to understand this issue to ensure that certain fundamental standards are followed so that all assessments administered to students are fair and consistent. This paper aims to explore the way students' writings are marked by the assessors at a selected university in Bangladesh. It addresses two questions: do all the markers follow the same criteria while marking a piece of writing? and, do test takers know the criteria used by the markers? For the current work, the variation of marks awarded by multiple markers on a written task was compared and the issues affecting their marking variations were explored. The data sample was chosen in simple random sampling approach to ensure representativeness of the population. The findings confirm no evidence of inter-marker reliability, where the marks of the script were clearly influenced by different factors for the individual markers. It also reveals that the test takers are unaware of the criteria used for marking their writing. The paper concludes by providing a number of recommendations on the way forward to solve the issues of fair and consistent assessment.

1. INTRODUCTION

The issues of fairness and consistency in assessment are a widespread concern for educators, students and parents. The core value of respect for the dignity and well-being of all students being assessed is to ensure fair and equitable assessment practices but unfortunately fairness and ethicality in educational tests and assessment have been at the center of the debates for long. Particularly in high-stakes contexts, it is clear that fairness should be a major concern to both the test developers, and to those being tested, as well to the assessors (Karami, 2018). An important element in addressing fairness issues in assessment is ensuring examination reliability, which is concerned with the consistency of test results by making sure that candidates obtain the result they deserve in a certain test, irrespective of who marks their paper, what types of questions are used, which topics are set or chosen to be answered, and when the examination is taken. The outcome of examination reliability is dependable, repeatable and consistent assessment score.

Inter-marker reliability is an important component of test reliability (Peter, 2017). In simple terms, inter-marker reliability means coming to an agreement about the quality of a particular student's work and reaching consensus about the marks or grades to be awarded to that work. Ensuring inter-marker reliability brings accountability to the assessment process; students are ensured fair assessment and are subject to the same expectations regardless of their section or instructor. The question of inter-marker reliability becomes crucial with the marking of subjective or more open-ended test items such as composition writing because, in such items, there are no single predefined correct answers to determine whether the answer displays the expected competence or knowledge of the test takers.

In Bangladeshi ELT context, it has been quite a common practice to mark or grade the subjective items with a holistic or an impressionistic approach. As a result, the marks or grade awarded on a student's response to any specific question or test by various evaluators have every chance to vary considerably to cause unreliability of the test, causing concerns not only to the student concerned but also to other stakeholders regarding their academic or career decisions. Current assessment policies and practices followed by ELT practitioners in Bangladesh in terms of ensuring fairness and equality is a research-worthy topic to uncover the real picture as well as to suggest strategies to ensure consistency and fairness.

2. THE STUDY

The current research aims to examine the existing practices in Bangladesh in assessing subjective written works of students in terms of reliability at tertiary level. It extends the current body of literature by combining empirical evidence of marking variance by multiple markers on the same script, reflective opinions on their principles of marking practice, and recommendations in ensuring inter-marker reliability. This study will, therefore, address the following research questions:

1. Do all the markers follow the same criteria while marking a piece of writing?
2. Do the test takers know the criteria used by the markers?

3. KEY ISSUES IN THE LITERATURE

Assessment in language education is the process of evaluating the performance of a person on a given task which can take different forms including tests, quizzes, interviews, written reports, remarks etc. to make inferences about their skill (Coombe, 2018). It is a systematic process of assessing and calculating collected data and information on the language comprehension, understanding and capacity of the students to enhance their language learning and progress. Starting from the second half of the 20th century, scholarly papers, textbooks, and course works on assessment across higher education institutions are contained in more numbers (Cheng, & Fox, 2017). The definition of fairness, in language assessment context, has been widely discussed since the late 1980s, but differences have been often found with regard to interpretation and context of the term (Kunnan, 2013). The word 'fair' is generally interpreted as in accordance with rules or standards. According to the Cambridge Dictionary (2020) 'fair' means to treat someone in a way that is right or reasonable, or to treat a group of people equally and not to allow personal opinions to influence one's judgement. Fairness in language assessment is used synonymously with equality with the goal of ensuring that all students have equal opportunities to represent what they know and what they can do (Turk, 2018). It is of utmost importance to ensure that all tests produce trustworthy results and are equal to any and all

test takers irrespective of their content areas. Any language assessment not fulfilling these essential prerequisites, is not reliable and valid, and thus inappropriate to use for all students.

While most of the linguists and assessment researchers traditionally coin the idea of fairness with equality in the assessment, Gipps & Stobart (2009) link the concept of fairness with equity to mean equality of opportunity i.e. access to similar resources and curricular opportunities, and state that, we will never achieve fair assessment, but we can always make it fairer than before. They plea for openness for the best protection against inequitable evaluation, noting that, openness should be ensured about design, constructs, assessment and grading of the test or assignment. This will bring out the principles and biases of the test design process to the test takers and provide an opportunity to discuss cultural and social factors and open up the relationship between evaluator and learner. Aitken (2012) articulated that linkages among student voice, assessment knowledge and pedagogical rationality contribute to fair student assessment practices.

There is a sustainable body of research defining the term reliability in language testing. Reliability is often defined as consistency of measurement. O'Mahony (2019) explains reliability by providing an example that, if students were to take a given test on a given day, and if their scores were similar to the scores they would have attained on another day irrespective of the markers, the test would be seen as highly reliable. On the other hand, if the scores were to differ wildly on different occasions, the reliability would be low. Therefore, a reliable test score is deemed to be consistent across different characteristics of the testing situation. Accordingly, reliability can be regarded as a function of the accuracy of scores from one set of tests to another. (Bachman, & Palmar, 1996; Loewenthal, & Lewis, 2018). Mehrens and Lehman (1987) describe reliability as the degree of consistency that occurs between two similar measures. Reliability is described to be the indicator of how stable, accurate, trustworthy and consistent a test is in every time testing the same thing (Worthen et al., 1993). McNamara (2000) defines reliability as "consistency of measurement of individuals by a test, usually expressed in a reliability coefficient" (p. 136). Weir (1988) notes that reliability as a basic requirement to be measured against any language test. A good number of studies argue that, reliability in assessment increases transparency, and decrease opportunities to insert any bias (Singh, 2014; Mohajan, 2017).

Usually, the assessment activities are divided into two types: formative and summative. Formative assessment means that the outcomes of an instructional initiative will be used in the creation and revision process (Dixson & Worrell, 2016). This type of assessment is used to improve educational initiatives, and it is the most common form of assessment in higher education constituting a large proportion of language learning assessment. The primary aim of formative assessment is to educate the student in a better way to enhance performance (Wiggins, 1998). On the other hand, summative assessment is cumulative assessment which is conducted to assess a student's learning or the quality of their learning, and judge their performance against some standards. Summative assessment extracts feedback to instructors about the quality of a subject or a program. Harlen & Gardner (2010) note that, in addition to its role in assessing a student's level of achievement or ability at a given time, summative assessment is often used to evaluate a student's eligibility for special programs such as gifted and talented education, to assess if a student can progress to the next grade level, to provide career guidance or to assess award qualifications.

Two marking models are widely used in assessing writing and speaking tasks, namely holistic marking and analytic marking (Khabbazbashi, & Galaczi, 2020). In short, holistic or cumulative scoring means giving a single overall score for the task as a whole; while analytic

scoring provides students with at least one rating score on each criterion, although the analytic scoring rubric also gives enough space on teachers to provide some input on each criterion.

There is a growing body of research literature centering on the issue of inter marker reliability that deal with various important aspects of the issue, such as impacts of rater's thinking process (Zhang, 2016) or the teacher's writing assessment literacy (Crusan et al, 2016), combination of holistic and analytic approaches (Tomas et al, 2019), impact of using objective rubric on assessing thesis (Williams & Kemp, 2019). However, to the best of the researcher's knowledge, none of the studies combined empirical evidence of heterogeneity of the markers in their marking practice with reflective views on their marking.

4. METHODS AND SAMPLING

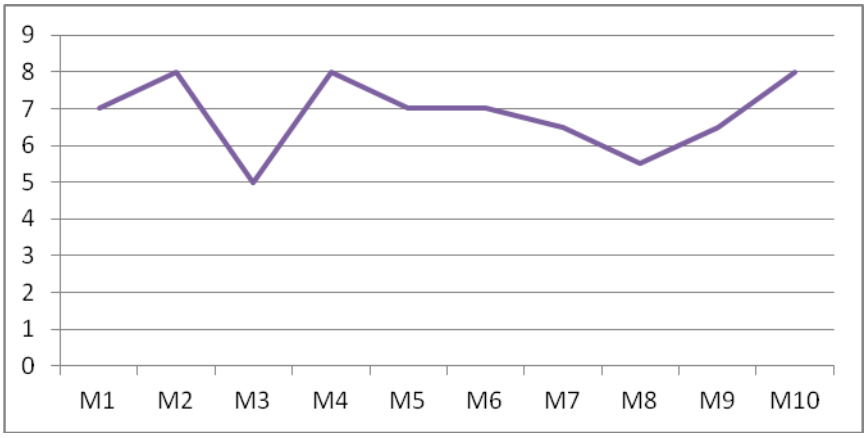
The study first compares marking on a same piece of writing task by ten markers and then records their principles of marking. The data sample was chosen through simple random sampling method to ensure representativeness of the population. Representativeness of a sample collected using simple random sampling makes it logical to generalize the results of the sample back to the population (Sharma, 2017). One participant was randomly selected out of 35 under-graduate students majoring in Pharmacy in a Bangladeshi university who were attending their EAP course in first semester. The participant was asked to write a paragraph of not more than 150 words on a given topic which served as the data. Ten markers from the Department of English of the same university were purposively selected based upon some criteria mentioned by Rai & Thapa (2015) which included their specialist knowledge, as well as, capacity and willingness to participate in the research. They were from English language or literature background with one to five years of teaching background at university level. The markers responded to a set of questions relating to their marking practices and beliefs when they completed marking.

The markers were also requested to make their recommendations for the actions necessary to ensure best assessment practice.

5. FINDINGS

5.1 Inconsistency in Marking

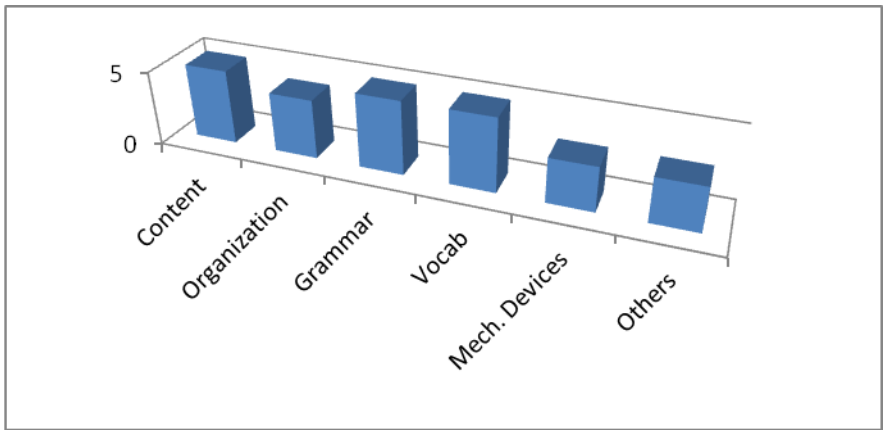
As the markers checked and marked the script in the first round, each of them followed a holistic scoring method by reading the whole piece and giving an overall score. The marks awarded by different markers varied significantly. The lowest mark given was 5 while 8 was awarded as the highest mark. The marks awarded by the 10 markers (M1 to M10) are shown in Figure 1 below.



[Figure 1: Marks awarded in the first round]

It is evident from the above figure that there was very little consistency among the markers while marking the script.

In the next phase of collecting data, the markers were individually asked to mention the criteria they had considered while marking the script. In response, they came up with a long list of all the possible factors that can come into play. However, the most commonly mentioned criteria by them were - content, organization, grammar, vocabulary, mechanical device and others, which included communication, cohesion and coherence etc. The findings of this phase are shown in Figure 2 below.



[Figure 2: Marking criteria mentioned by the markers]

5.2 Markers' Views on their Grading

The ten selected markers were asked to give their views on various aspects of grading they usually did. The following responses were generated.

While responding to the question, "Do you follow a consistent method of marking rubric in marking students' works?", three out of ten markers said 'yes' and 4 said 'not always'. Only two markers said 'no' to this question and one did not respond.

As the markers were asked, "Do you have a policy to share your marking/grading principles with your students?", all of the markers said, "no" to this question stating that there is no policy of sharing marking principles with the students.

In their response to the question 'Do you talk to a fellow marker about the marking criteria before or after marking?' four out of the ten markers said 'yes', two said 'sometimes' and three said 'no'.

When markers were asked if they thought that their students were happy with the marks given, six markers out of ten said 'yes'. However, three markers said 'no' and two said, 'not always'.

As the markers were asked whether they had enough time for doing the marking, eight out of ten markers said 'yes' and only two said 'no'.

While responding to the question, "Is there any provision of double/second marking in your institution?" all the ten markers said 'no'. However, seven out of ten markers believe that double/second marking could produce more reliable results though two markers think that double/second marking may not always ensure more reliable test results.

It is also revealed from the study that six markers out of ten have no training on language testing and assessment. Though the selected markers were all chosen from one reputed university located in Dhaka metropolitan area, only four of them had received training on language testing.

6. DISCUSSIONS

The survey data collected on individual markings in the first round indicate remarkable inconsistency among the markers, which imply marking unreliability of the assessment. It also became apparent that individual consideration of marking factors among the markers differ considerably from one another. The findings also report that the markers do not follow any identical marking rubric. A marking rubric describes the criteria and marks available for aspects of a task, assignment or examination script (Courtney, 2020). For example, the aspects in a written task can be task fulfillment, organization and creativity, linguistic skills etc. Numerous studies argue that an effective use of marking rubric is one of the basic principles to ensure fairness in assessment (Atkinson & Lim, 2013; Leader & Clinton, 2018; Grainger & Weir, 2020).

The findings reveal that the participating markers follow certain self-determined criteria in marking their students' tasks which are not similar to each other. This study also finds that the markers do not have a policy of sharing their marking principles with their students. Therefore, the students cannot see the allocation of marks with performance levels or indices in advance. This finding challenges research-established policy of fair assessment that identifies a linkage between student voice with assessment knowledge (Atkin, 2012).

The data also unfold that there is no practice of discussion among the fellow markers about their individual marking criteria before or after marking. However, numerous previous studies including Baird et al (2010) promote the beneficial effects of discussion of the marking scheme among the fellow markers in ensuring inter-marker reliability. The collected data also indicate that there is no provision of double/second marking in the research site institution, which also puts fair and reliable marking in question (Bloxham et al, 2016).

7. KEY RECOMMENDATIONS

Based on the above findings, it is undeniable that appropriate measures are strictly needed to ensure consistent, correct and fair marking practice by the single marker as well as across different markers. The following recommendations are proposed from the summary of suggestions made by the ten participating markers to way forward:

a) An analytical marking scheme with several parameters may be used to assess writing activities based on different aspects of writing skills such as content, structure, usage of languages, vocabulary etc. This can ensure reliability and consistency of marking scripts across markers. However, in case of holistic marking, at least a brief description of the various grades to be achieved should be defined. It may be noted that numerous preceding studies prefer analytical marking scheme over holistic marking scheme to ensure inter-marker reliability (Wright & Masters, 1982; Urbach, 2014; Rios et al, 2017).

b) Markers may be trained on applying rubric while marking. Developing and sharing a rubric for an assignment among markers and students before and after the marking provides an overview of what an "A+" assignment will look like. A well-developed rubric can also act as a measuring stick for the marker when marking where student works are rated according to the guidelines. A large number of prior studies strongly

corroborate this recommendation (Atkinson & Lim, 2013; Edwards, 2017; Leader & Clinton, 2018; Courtney, 2020; Grainger & Weir, 2020).

c) Pre-marking and post-marking discussion among the markers is helpful for a clearer idea about what the grades or marks to be awarded for what quality of work, especially when there is no rubric for the markers to guide. Therefore, making expectations clear and sticking to them is the key. This factor has, however, been established by many earlier studies towards ensuring consistent and fair assessment (Hume & Coll, 2009; Hermansen, 2014).

d) The markers may grade assignments in groups, if necessary. Spending some time marking assignments together can help to make sure the markers are on the same page as their colleagues. This method has also been found to be successful in previous studies to ensure assessment consistency and integrity (Bird & Yucel, 2013).

e) To ensure accuracy in assessment in a more structured process, there could be provisions of double/second marking, and moderation. Effectiveness of this step in achieving reliability among the markers has been verified by many previous researches (Chen et al, 2017; Burger, 2017).

To summarize the recommendations, fair and consistent grading starts with setting specific criteria for the expected quality of work. It also includes establishing practices, follows from communicating such expectations and practices with colleagues and students, and ends with engaging in these practices throughout the assessment process.

8. CONCLUSION

This study attempted to identify the current practices in assessing students' writings at a university in Bangladesh by comparing marks awarded by multiple markers to a selected sample of writing. It also integrated views of the markers concerned to investigate issues related to the variations in their markings. Although there is no denying that each student should be awarded the mark that they actually deserve, it has always been a challenging task to ensure fairness in marking students' works, especially, the free or more open-ended writings tasks. Therefore, no one can rely on holistic or impressionistic marking which does not produce consistent test scores as evident from this study. It is strongly believed that the markers and test developers could resolve the issue by acting on the set of recommendations emerged from this study.

REFERENCES

Aitken, N., Webber, C. F., Lupart, J., Scott, S., & Runté, R. (2011). Assessment in Alberta: Six areas of concern. *The Educational Forum*, 75, 192–209.

Atkinson, D., and S. L. Lim. 2013. "Improving Assessment Processes in Higher Education: Student and Teacher Perceptions of the Effectiveness of a Rubric Embedded in a LMS." Australasian Journal of Educational Technology 29 (5): 651–666.

Baird, J. A., Greatorex, J., & Bell, J. F. (2010). What makes marking reliable? Experiments with UK examinations. *Assessment in education: Principles, policy & practice*, *11*(3), 331-348.

Bachman, L.F. & A.S. Palmar. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford, Oxford University Press.

Bird, F. L., & Yucel, R. (2013). Improving marking reliability of scientific writing with the Developing Understanding of Assessment for Learning programme. *Assessment & Evaluation in Higher Education*, *38*(5), 536-553.

Bloxham, S., Hughes, C., & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment & Evaluation in Higher Education*, *41*(4), 638-653.

Burger, R. (2017). Student perceptions of the fairness of grading procedures: a multilevel investigation of the role of the academic environment. *Higher Education*, *74*(2), 301-320.

Chen, C. Y., Chang, H., Hsu, W. C., & Sheen, G. J. (2017). Learning, behaviour and reaction framework: a model for training raters to improve assessment quality. *Assessment & Evaluation in Higher Education*, *42*(5), 705-723.

Cheng, L., & Fox, J. (2017). Assessment in the language classroom: Teachers supporting student learning. London, England: Palgrave. 200 pp, paperback.

Coombe, C. (2018). *An A to Z of Second Language Assessment: How Language Teachers Understand Assessment Concepts.* London, UK: British Council.

Courtney, J. (2020) Marking Rubrics, Learning, Teaching & Technology Centre, Technological University Dublin.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing writing*, *28*, 43-56.

Dictionary, C. (2020). Cambridge Online Dictionary. https://dictionary.cambridge.org/dictionary/english/fair. accessed on April 8, 2020

Dixson, D. D., & Worrell, F. C. (2016). Formative and summative assessment in the classroom. *Theory into practice*, *55*(2), 153-159.

Edwards, F. (2017). A rubric to track the development of secondary pre-service and novice teachers' summative assessment literacy. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 205-227.

Grainger, P., & Weir, K. (Eds.). (2020). *Facilitating Student Learning and Engagement in Higher Education through Assessment Rubrics*. Cambridge Scholars Publishing.

Gipps C., Stobart G. (2009) Fairness in Assessment. doi.org/10.1007/978-1-4020-9964-9_6. In: Wyatt-Smith C., Cumming J.J. (eds) Educational Assessment in the 21st Century. Springer, Dordrecht.

Harlen, W., & Gardner, J. (2010). Assessment to support learning. In J. Gardner, W. Harlen, L. Hayward, G. Stobart, & M. Montgomery (Eds.), *Developing teacher assessment* (pp. 15–28). New York, NY: Open University Press.

Hermansen, H. (2014). Recontextualising assessment resources for use in local settings: Opening up the black box of teachers' knowledge work. *Curriculum Journal*, *25*(4), 470-494.

Hume, A., & Coll, R. K. (2009). Assessment of learning, for learning, and as learning: New Zealand case studies. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 269-290.

Karami, H. (Ed.). (2018). *Fairness issues in educational assessment*. Routledge.

Khabbazbashi, N., & Galaczi, E. D. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment. *Language Testing*.

Kunnan, A. J. (2013). Fairness and justice in language assessment. *The companion to language assessment*, *3*, 1098-1114.

McNamara, T. (2000). *Language testing.* Oxford: Oxford University Press.

Mehrens, W. A. & Lehmann, I. J. (1987). Using standardized tests in education. New York: Longman.

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University. Economic Series*, *17*(4), 59-82.

Leader, D. C., & Clinton, M. S. (2018). Student perceptions of the effectiveness of rubrics. *Journal of Business and Educational Leadership*, *8*(1), 86-103.

Loewenthal, K. M., & Lewis, C. A. (2018). *An introduction to psychological tests and scales*. Psychology press.

O'Mahony, C. (2019). Reliability and Validity in Language Testing–A Real Conflict? *英語教育研究所紀要 (CELE Journal)= CELE JOURNAL*, (27), 133-140.

Peter, C. (2017). Importance of Standards in Testing Technical English for Engineering Students of Tamil Nadu. *International Journal of Educational Sciences*, *16*(1-3), 36-42.

Pettifor, J. L., & Saklofske, D. H. (2012). Fair and ethical student assessment practices. In C. F. Webber & J. Lupart (Eds.), *Leading student assessment* (pp. 87–106). Dordrecht: Springer.

Rai, N., & Thapa, B. (2015). A study on purposive sampling method in research. *Kathmandu: Kathmandu School of Law*.

Rios, J. A., Sparks, J. R., Zhang, M., & Liu, O. L. (2017). Development and validation of the written communication assessment of the HEIghten® Outcomes Assessment Suite. *ETS Research Report Series*, *2017*(1), 1-16.

Sharma, G. (2017). Pros and cons of different sampling techniques. *International journal of applied research*, *3*(7), 749-752.

Shelleyann Scott, Charles F. Webber, Judy L. Lupart, Nola Aitken & Donald E. Scott (2014) Fair and equitable assessment practices for all students, *Assessment in Education: Principles, Policy & Practice*, 21:1, 52-70, DOI: 10.1080/0969594X.2013.776943

Singh, A. S. (2014). Conducting Case Study Research in Non-Profit Organisations. Qualitative Market Research: An International Journal, 17,77–84.

Tomas, C., Whitt, E., Lavelle-Hill, R., & Severn, K. (2019). Modelling holistic marks with analytic rubrics. In *Frontiers in Education* (Vol. 4, p. 89). Frontiers.

Turk, H. (2018). What is Fair? Case Study and Analysis of Second Language Acquisition and Assessment.

Urbach, D. (2014). Examining the factor structure of writing assessment based on sets of analytical marking criteria. *Procedia-Social and Behavioral Sciences*, *141*, 1106-1111.

Weir, C. J. (1998). *Communicative Language Testing: With Special Reference to English as a Foreign Language*. Exeter: Exeter University.

Wiggins, G. (1998). *Educative Assessment. Designing Assessments To Inform and Improve Student Performance*. Jossey-Bass Publishers, 350 Sansome Street, San Francisco, CA 94104.

Williams, L., & Kemp, S. (2019). Independent markers of master's theses show low levels of agreement. *Assessment & Evaluation in Higher Education*, *44*(5), 764-771.

Worthen, B. R., Borg, W. R., and White, K. R. (1993). Measurement and evaluation in the school. NY: Longman.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, *27*, 37-53.

Author biodata:

Takad Ahmed Chowdhury ***(ORCID ID: https://orcid.org/0000-0002-0785) is a Ph.D. candidate at the School of Educational Studies, Universiti Sains Malaysia. He is also an associate professor at the University of Asia Pacific, Bangladesh. His research interests include ESP/EAP, ESL/EFL writing, curriculum development, and English literature.***